Essays in robust estimation and inference in semi- and nonparametric econometrics

Yannick Guyonvarch

université PARIS-SACLAY

ENSAE

IP PARIS

NNT : 2019SACLG007

Thèse de doctorat

Contributions à l'estimation et à l'inférence robuste en économétrie semi- et nonparamétrique

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Ecole nationale de la statistique et de l'administration économique (ENSAE Paris)

Ecole doctorale n°578 - École doctorale: Science de l'Homme et de la Société (SHS)
Spécialité de doctorat: Sciences économiques

Thèse présentée et soutenue à Palaiseau, le 28 novembre 2019, par

YANNICK GUYONVARCH

Composition du Jury :

Xavier d'Haultfœuille Professeur, CREST-ENSAE	Invité
José de Sousa Professeur, Université Paris-Sud	Président du jury
Pascal Lavergne Professeur, Toulouse School of Economics	Rapporteur
Anna Simoni Professeur, CREST-ENSAE	Directrice de thèse
Martin Weidner Professeur, University College London	Rapporteur

Remerciements

Malgré l'unique nom inscrit sur la couverture de ce manuscrit, la thèse est une expérience intrinsèquement collective. Les quelques paragraphes qui suivent visent à remercier de manière aussi exhaustive que possible les personnes qui, de près ou de loin, ont eu une influence sur l'aboutissement de ce travail.

Je tiens tout d'abord à remercier Anna Simoni de m'avoir donné l'opportunité de réaliser cette thèse, ainsi que José de Sousa, Pascal Lavergne et Martin Weidner de faire partie de mon jury.

Il y a maintenant presque 5 ans, j'ai croisé pour la première fois la route de Xavier D'Haultfœuille. J'étais alors élève en deuxième année à l'ENSAE. Son cours d'économétrie et sa compréhension profonde de cette discipline m'ont marqué. Il m'a ensuite accueilli au CREST en stage dès l'été 2015. Dès le début de ce stage, lui et son co-auteur Clément de Chaisemartin m'ont fait confiance: ils m'ont lancé sur un projet ambitieux où je n'avais pas forcément toutes les compétences requises. J'ai beaucoup appris au cours de ce stage et cela m'a définitivement convaincu de faire une thèse. Xavier m'a ensuite suivi tout au long de la troisième d'année sur mon mémoire de recherche sur un sujet qui était assez loin de ses centres d'intérêt habituels et sur lequel je n'avais une fois encore pas forcément le profil type. Là aussi, il m'a fait confiance. Ces deux expériences ont finalement conduit aujourd'hui à plusieurs projets aboutis ou en bonne voie dont deux sont présents dans cette thèse. Pour toutes ces raisons et bien d'autres, je lui suis extrêmement reconnaissant.

Je dois aussi rendre un hommage appuyé à Laurent Davezies. Tout comme Xavier, il possède une compréhension rare de nombreux aspects fondamentaux de l'économétrie. Je ne saurais compter le nombre d'heures où nous avons discuté de problèmes économétriques dont une part non négligeable ne dépassera vraisemblablement jamais les portes du CREST!

Un très grand merci également à Alexis Derumigny et Lucas Girard qui ont directement contribué à ce manuscrit et avec qui j'espère continuer à travailler et interagir longtemps.

La liste des portes du CREST auxquelles je suis venu frapper avec des questions est longue mais deux personnes y ont particulièrement été exposées: je salue ici Guillaume Lecué qui m'a beaucoup éclairé sur nombre de questions de processus empiriques et Christophe Gaillac avec qui j'ai beaucoup échangé (et cela va sans nul doute continuer!) sur les mystères des problèmes inverses et des conditions de source.

Ces trois dernières années, le CREST a déménagé, s'est agrandi mais une chose n'a pas changé: c'est un endroit chaleureux. J'ai eu la chance de côtoyer de nombreuses générations de doctorants et assistants de recherche: Marianne Bléhaut, Julie Pernaudet, Manon Garrouste, Pauline Rossi, Meryam Zaiem, Daphné Skandalis, Alicia Marguerie, Malka Guillot, Jérôme Trinh, Ivan Ouss, Benjamin Walter,

Audrey Rain, Sandra Nevoux et Victor Lyonnet qui m'ont accueilli au début; le labo d'économie et mes partenaires de l'équipe de foot du Mancrester United qui compte dans ses rangs le valeureux et déjà cité Lucas Girard (capitaine et milieu récupérateur), Gwen-Jiro Clochard, Fabinho Perez, Guido Bodrato et Solide Tang (défenseurs), Rémi Box-to-box, Aurélien Bigo, Thomas Delemotte, Optimus Mugnier, Julien Monardo, Hugo Molina, Antoine Valtat, Thibault Cézanne et Ahmed Diop (latéraux), Reda Aboutajdine, Etienne Giggs, Morgane Cure dite MC9, Morgane Guignard, Anasuya Raj, Bérengère Patault, Clémence Lenoir, Big J, Sophie Nottermeyer et Alice Lapeyre (attaquants) et Germain Gauthier, Elio Nimier-David dit le Normand, Pauline Carry, Léa Bou Sleiman, Elia Pérennès, Esther Mbih, Emilie Sartre, Maxime Cugnon de Sévricourt, Heloïse Clolerly, Lorenzo Kaaks, Clémence Tricaud, Pierre-Edouard Collignon, Antoine Ferey, Antoine Bertheau, Roxane Morel, Ao Wang, Arne Uhlendorff, Benoît Schmutz, Christophe Bellego, David Benattia, Roxana Fernandez, Franck Malherbet, Isabelle Méjean, Thibaud Vergé, Philippe Choné, Pierre Boyer, Laurent Linnemer, Thierry Kamionka (supporters) et Francis Kramarz (président du club); le labo de statistique et son équipe de foot du Crestiano composés de Bad Badr (capitaine), Arnak Dalalyan et Alexander Buchholz (défenseurs), Lionel Riou-Durand, Geoffrey Chinot et Aurélien Brouillaud (latéraux), Mohamed Ndaoud, Léna Carel, Solenne Gaucher (attaquants), Jules Depersin (manager) et Lucie Neirac, Avo Karagulyan, Amir-Hossein Bateni, Nicolas Schreuder, François-Pierre Paty, Gabriel Ducrocq, Boris Muzellec, Victor-Emmanuel Brunel, Cristina Butucea, Vincent Cottet, Nicolas Chopin, Pierre Alquier, Marco Cuturi et Sacha Tsybakov (supporters); Jérémy "the Beast" L'Hour qui est un économiste infiltré au labo de statistique ou l'inverse.

Un laboratoire de recherche a bien évidemment besoin de nombreux métiers, au-delà des seuls chercheurs, pour fonctionner. Je tiens à saluer ici Pascale Deniau, Edith Verger, Murielle Jules, Fanda Traore, Marie-Christine Baker, Michèle Arger, ainsi que Uolo, Julia et Antonia dont je croise la route tôt chaque matin.

Je remercie enfin ma famille, mes amis d'Orsay, de l'université Paris-Sud, de l'ENSAE et de mon club de volley de Villebon-sur-Yvette.

Contents

Chapter 1

Introduction/ Résumé substantiel en français

Quelle est la connexion entre ce manuscrit de thèse et des problématiques économiques concrètes? Une telle question semble légitime au vu du titre quelque peu technique du travail présenté ici. Le paragraphe introductif qui suit aborde principalement ce sujet.

Comment est-ce que les ménages à bas revenus répartissent leur budget entre des biens luxueux - tels que les vêtements de marque - et les dépenses plus fondamentales comme la nourriture? Les employés au chômage ont-ils de plus grandes chances de retrouver un emploi quand ils participent à une formation auprès de Pôle Emploi? Ces deux questions sont des exemples de sujets qui intéressent les économistes ([46, 47]). Dans le premier exemple, le but est de comprendre les mécanismes derrière les décisions économiques au niveau individuel. Dans le second cas, l'intérêt réside principalement dans l'évaluation de l'impact d'une politique publique, ici de retour à l'emploi. Pour répondre à ces questions, la théorie économique fournit des prédictions qui doivent être testées à partir de données réelles. Pour tester des prédictions économiques, des restrictions doivent être imposées sur la façon dont les données sont engendrées. Ces contraintes forment un modèle des comportements observés. Il est peu plausible d'affirmer que nous pouvons expliquer parfaitement la consommation (respectivement le retour à l'emploi) en fonction du revenu alloué par les ménages (respectivement des dépenses de formation). Il est plus raisonnable de supposer que la consommation ou le retour à l'emploi dépendent également de facteurs inobservables dans les données qui capturent des mécanismes complexes et indicibles. Quand les composantes observées et inobservées du modèle sont traitées comme aléatoires, nous obtenons un modèle statistique. Dans la veine du chapitre introductif de [60], les statistiques peuvent être décrites comme l'interface générique entre des théories que nous cherchons à tester et des données. Notre travail se rattache à une discipline appelée économétrie. Cette dernière est un sous-champ de l'économie qui utilise des outils statistiques pour répondre à des questions socio-économiques. Dans cette introduction, nous cherchons à comparer l'économétrie avec plusieurs sous-champs des statistiques et en particulier l'apprentissage statistique. L'apprentissage statistique est une discipline qui étudie les propriétés théoriques d'algorithmes d'apprentissage automatique (machine learning en anglais) quand les données sont supposées être générées selon un modèle statistique. Comme nous le soulignons plus bas, l'économétrie et l'apprentissage statistique diffèrent dans leur définition d'un modèle statistique. Remarquons par ailleurs que selon le niveau de généralité du modèle statistique considéré, celui-ci sera appelé paramétrique, semiparamétrique ou nonparamétrique. Nous donnons des définitions précises de ces notions dans le reste de l'introduction.

Notions clés en statistiques semi- et nonparamétriques

Ce manuscrit se concentre sur les modèles statistiques dits semi- et nonparamétriques. Pour expliquer précisément ces notions et mieux les comprendre, quelques définitions sont de rigueur. Nous considérons un vecteur aléatoire W qui va d'un espace probabilisé sous-jacent $(\Omega, \mathcal{A}, P)$ vers un espace mesurable $(E, \mathcal{E})$. W fait référence à toutes les composantes aléatoires du modèle, quelles soient observables ou non. Nous supposons que E peut être muni d'une structure d'espace métrique grâce à la norme $\|\cdot\|_E$. $\mathcal{Q}$ correspond à l'ensemble des lois de probabilité définies sur $(E, \mathcal{E})$. Dans ce travail, nous considérons toujours que la loi de W dénotée Q_W appartient à un sous-ensemble strict de $\mathcal{Q}$ que nous appelons $\mathcal{Q}^*$. Un exemple classique est $\mathcal{Q}^* := \left\{Q \in \mathcal{Q} : \mathbb{E}_Q\left[||W||_E^2\right] < +\infty\right\}$, où $\mathbb{E}_Q$ désigne l'opérateur d'espérance sous la loi Q. Cet exemple est un sous-ensemble nonparamétrique de $\mathcal{Q}$ car les éléments dans $\mathcal{Q}^*$ ne sont pas pleinement caractérisés par un paramètre fini-dimensionnel. Nous allons en fait nous intéresser uniquement à des sous-ensembles nonparamétriques de $\mathcal{Q}$ ici. Un modèle statistique est construit en: i) choisissant un ensemble Θ appelé l'ensemble des paramètres; ii) en associant à chaque $\theta \in \Theta$ une distribution $Q_\theta \in \mathcal{Q}^*$. Pour fixer les idées, nous donnons l'exemple du modèle canonique de régression linéaire en nous inspirant du chapitre introductif de [60]: $Z_o = Z_e'\beta + \epsilon$, avec $Z_e \in \mathbb{R}^p$. Les notations inhabituelles Z_o pour la variable expliquée et Z_e pour le vecteur de variables explicatives sont introduites par souci de cohérence avec les chapitres suivants. Nous écrivons $Z = (Z_o, Z_e')'$. Dans cet exemple, le paramètre est $\theta = (\beta, Q_{Z_e,\epsilon})$. L'ensemble des paramètres est $\Theta = \mathbb{R}^p \times \mathcal{D}$ avec $\mathcal{D} := \left\{Q : \mathbb{E}_Q[Z_e\epsilon] = 0, \mathbb{E}_Q[Z_eZ_e']^{-1} < +\infty\right\}$ et $\mathcal{Q}^* = \left\{Q : \mathbb{E}_Q[Z_eZ_e']^{-1} < +\infty\right\}$. Dans le modèle de régression linéaire, nous nous intéressons uniquement à β qui peut s'écrire formellement $\beta = T(\theta)$ pour T une projection. Il est fréquent que le paramètre d'intérêt ne soit pas θ lui-même mais une transformation de celui-ci. Lorsque $T(\theta)$ est une quantité fini-dimensionnelle, nous appelons le modèle semiparamétrique, sinon nous parlons de modèle nonparamétrique.

La question de l'identification d'un modèle statistique est fondamentale: un modèle est dit identifié si tout $Q \in \mathcal{Q}^*$ peut être généré par au plus un $\theta \in \Theta$. Dans ce qui suit, nous faisons l'hypothèse que le modèle est identifié. Donner des conditions suffisantes d'identification n'est pas chose facile en général et sort du cadre de ce manuscrit. Il faut néanmoins garder à l'esprit que nous nous focalisons sur des modèles pour lesquels la question de l'identification est (plutôt) bien comprise. Dans le cas du modèle linéaire, les restrictions $\mathbb{E}_Q[Z_eZ_e']^{-1} < +\infty$ et $\mathbb{E}_Q[Z_e\epsilon] = 0$ sont nécessaires et suffisantes pour l'identification par exemple.

Quand un modèle est identifié, ses paramètres peuvent être exprimés en fonction de la distribution des variables aléatoires observées ([60]). Etant donnée l'identification du modèle, la principale tâche d'un économètre est d'utiliser des observations pour estimer $T(\theta)$ et faire de l'inférence sur cette quantité. A partir de maintenant, nous supposons que nous avons à notre disposition n observations $(Z_i)_{i=1}^n$ de la loi jointe Q_n. Nous imposons aussi que les observations aient toutes la même loi marginale, *i.e* soient identiquement distribuées. Nous restreignons aussi le domaine de définition de $T(\theta)$: ce-dernier appartient à un espace métrique $(\mathcal{T}, \|\cdot\|_T)$. Un estimateur est une fonction mesurable de $(Z_i)_{i=1}^n$ qui prend ses valeurs dans $\mathcal{T}$. La qualité d'un estimateur est mesurée par $\left|\left|\widehat{T}(\theta) - T(\theta)\right|\right|_T$. Un estimateur est convergent si $\left|\left|\widehat{T}(\theta) - T(\theta)\right|\right|_T$ tend vers 0 lorsque n augmente. En général, le choix de la norme $\|\cdot\|_T$ n'est pas unique. Quand $\mathcal{T}$ est fini-dimensionnel, ce choix n'est pas crucial vu que toutes les normes sont alors équivalentes. D'un autre côté, lorsque $\mathcal{T}$ est de dimension infinie, les normes ne sont plus toutes équivalentes. Il se peut qu'alors un estimateur soit convergent pour une norme mais pas pour une autre. En dimension infinie, la différence entre différentes normes peut s'avérer très utile: il est parfois possible d'utiliser une norme comme un outil de régularisation pour faciliter la convergence d'un estimateur par rapport à une autre norme. La notion de régularisation statistique est expliquée

plus en détails après. L'inférence regroupe deux sujets étroitement liés: les intervalles de confiance et les tests d'hypothèses. Comme nous ne nous intéressons qu'aux intervalles de confiance dans les chapitres ultérieurs, nous laissons de côté la définition des tests d'hypothèses ici. Ce qui suit s'appuie essentiellement sur le chapitre 6 de [79]. En quelques mots, un ensemble de confiance (EC) est un sous-ensemble aléatoire C_n de $\mathcal{T}$ qui dépend de $(Z_i)_{i=1}^n$ mais pas de $T(\theta)$. Nous présentons maintenant les critères asymptotiques qui sont communément admis pour évaluer la qualité d'un EC. Etant donné $\delta \in (0,1)$, un EC est de niveau asymptotique $1-\delta$ ponctuellement sur Θ si

$$\inf_{\theta\in\Theta} \liminf_{n\to+\infty} \mathbb{P}_{Q_{\theta,n}}(C_n \ni T(\theta)) \geq 1-\delta, \tag{1.1}$$

et il est de niveau asymptotique $1-\delta$ uniformément sur Θ si

$$\liminf_{n\to+\infty} \inf_{\theta\in\Theta} \mathbb{P}_{Q_{\theta,n}}(C_n \ni T(\theta)) \geq 1-\delta. \tag{1.2}$$

Le deuxième critère ([104, 29]), qui est parfois appelé le critère d'honnêteté, est de toute évidence plus exigeant que le premier et il a été beaucoup étudié, en particulier en statistiques nonparamétriques. Ces deux premiers critères assurent que l'EC est fiable asymptotiquement. Ils ne sont cependant pas suffisants car ils n'excluent pas des EC triviaux: rien n'empêche avec les deux précédents critères de prendre $C_n = \mathcal{T}$ pour tout n. Nous exigeons donc également d'un EC qu'il soit optimal dans un certain sens. L'optimalité peut être définie de plusieurs manières. Un EC peut être dit ponctuellement/uniformément optimal si l'inégalité dans (1.1)/(1.2) devient une égalité. Une autre règle communément employée requiert que le diamètre de l'EC décroisse vers zéro en probabilité suffisamment vite quand n tend vers l'infini. Pour choisir entre deux ECs qui vérifient les critères d'optimalité ci-avant, il est possible d'étudier la limite du ratio des diamètres des deux ECs.

Le paradigme de la minimisation du risque empirique régularisé (MRER) en apprentissage statistique

Pour discerner les connexions et les différences entre l'économétrie et l'apprentissage statistique, nous devons tout d'abord comprendre le but général de l'apprentissage statistique et le cadre théorique qui en découle. Nous nous concentrons sur le cas où les données observées peuvent être divisées en deux: une variable à prédire $Z_o \in \mathcal{Z}_o \subseteq \mathbb{R}$ et un ensemble de prédicteurs potentiels $Z_e \in \mathcal{Z}_e$, avec une distribution jointe Q_{Z_o,Z_e}. Nous gardons la notation $Z = (Z_o, Z_e')'$. Le but est de prédire Z_o aussi précisément que possible à l'aide d'une fonction de Z_e, selon une règle qui définit la qualité de la prédiction. La règle de prédiction (également appelée perte) et la classe de fonctions sont choisies par le statisticien et ces choix sont grandement motivés par des considérations computationnelles. De façon formalisée, le problème théorique est: étant données une classe de fonctions $\mathcal{H}$ allant de $\mathcal{Z}_e$ dans $\mathcal{V}$, et une perte $\ell : \mathcal{Z}_o \times \mathcal{Z}_e \times \mathcal{H} \to \mathbb{R}^+$, nous faisons l'hypothèse qu'il existe $h^* \in \mathcal{H}$ non nécessairement unique tel que

$$h^* \in \operatorname*{argmin}_{h\in\mathcal{H}} \mathbb{E}_{Q_Z}\left[\ell(Z_o, Z_e, h)\right]. \tag{1.3}$$

Par exemple, nous pourrions prendre $\ell(z_o, z_e, h) = (z_o - h(z_e))^2$, *i.e* la perte des moindres carrés, et $\mathcal{H} := \left\{h : \sup_{z_e\in\mathcal{Z}_e} |h(z_e)| \leq M\right\}$. Cela revient à résoudre un problème de moindres carrés nonparamétriques. Remarquons qu'en économétrie, nous ne nous intéressons aux moindres carrés nonparamétriques que si nous faisons l'hypothèse que les données sont générées selon le modèle $Z_o = h^{**}(Z_e) + \epsilon$, sous la contrainte $\mathbb{E}_{Q_{\epsilon|Z_e}}[\epsilon \mid Z_e] = 0$. En effet, un résultat classique montre que si $h^{**} \in \mathcal{H}$, alors h^{**} satisfait (1.3). Que se passe-t-il quand $h^{**} \notin \mathcal{H}$? Le problème (1.3) est toujours bien défini et admet une solution mais cette solution n'est pas h^{**} et est sous-optimale: nous avons $\mathbb{E}_{Q_Z}[\ell(Z_o, Z_e, h^{**})] < \mathbb{E}_{Q_Z}[\ell(Z_o, Z_e, h^*)]$. En termes économétriques, résoudre (1.3) quand $h^{**} \notin \mathcal{H}$ est équivalent à s'intéresser à un modèle

mal spécifié. Dans le cadre de l'apprentissage statistique, les modèles mal spécifiés sont généralement autorisés. Le paradigme de l'apprentissage statistique a d'autres particularités: l'intérêt est principalement porté sur des modèles dits de grande dimension dans lesquels la classe de fonctions $\mathcal{H}$ peut croître avec n. Dans un modèle typique de grande dimension, $h(Z_e)$ prend la forme $Z_e'\beta$ où $\beta \in \mathbb{R}^p$ et p est potentiellement beaucoup plus grand que n. Pour rendre le problème solvable, une hypothèse classique est celle de sparsité, *i.e* seulement s entrées (avec s petit par rapport à n) sont non-nulles dans le vecteur β. Une généralisation de la sparsité appelée sparsité approximative est aussi courante: elle impose que β soit bien approximé (et non plus exactement déterminé) par un faible nombre d'entrées. La sparsité approximative entretient des liens étroits avec les modèles nonparamétriques classiques et est proche de la notion de régularité d'une fonction. Nous renvoyons le lecteur vers [15], [16] et [17] pour des discussions éclairantes sur le sujet. Comment h^* peut-elle être reconstruite à partir d'observations? Les n observations dans l'échantillon $(Z_i)_{i=1}^n$ sont supposées indépendantes et identiquement distribuées (*i.i.d*) et une approche naïve consisterait à prendre directement la contrepartie empirique de (1.3): $h_n \in \operatorname{argmin}_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \ell(Z_{o,i}, Z_{e,i}, h)$. Ceci n'est cependant pas satisfaisant dans un cadre de grande dimension. En fait, le problème de minimisation empirique que nous venons de définir ne tire aucunement parti de l'hypothèse de sparsité. Pour y remédier, il faut doter $\mathcal{H}$ d'une norme $\|\cdot\|_R$ qui capte bien la notion de sparsité et utiliser cette norme pour régulariser la procédure de minimisation empirique. Le problème devient

$$h_n \in \operatorname*{argmin}_{h\in\mathcal{H}} \left\{ \frac{1}{n}\sum_{i=1}^n \ell(Z_{o,i}, Z_{e,i}, h) + \alpha\, ||h||_R^p \right\}, \tag{1.4}$$

et est appelé la procédure de minimisation du risque empirique régularisé (MRER). La quantité α est le poids associé au terme de régularisation et son choix est clé pour obtenir des garanties théoriques sur h_n. L'exposant p est le plus souvent choisi égal à 1 ou 2. La procédure la plus connue qui s'inscrit dans ce cadre général est le Lasso ([133]) pour la régression linéaire: $\ell(Z_{o,i}, Z_{e,i}, h) = (Z_{o,i} - Z_{e,i}'\beta)^2$, $\mathcal{H} = \{\langle\cdot, \beta\rangle, \beta \in \mathbb{R}^p\}$, $p = 1$ et $||h||_R = ||\beta||_1$ avec $||\cdot||_1$ la norme ℓ_1 dans $\mathbb{R}^p$. La qualité du minimiseur estimé h_n est mesurée par le critère dit de l'excès de risque $\mathcal{R}(h_n) := \mathbb{E}_{Q_Z}\left[\ell(Z_o, Z_e, h_n) \mid (Z_i)_{i=1}^n\right] - \mathbb{E}_{Q_Z}\left[\ell(Z_o, Z_e, h^*)\right]$. Dans la définition précédente, (Z_o, Z_e) est une copie indépendante de la suite $(Z_{o,i}, Z_{e,i})_{i=1}^n$. En apprentissage statistique, le but est de contrôler la probabilité que l'excès de risque soit plus grand qu'un seuil explicite pour un nombre donné d'observations. Ceci s'appelle une inégalité oracle et sa forme générale est la suivante: pour tout $\delta \in (0,1)$ et tout $n \geq 1$, $\mathbb{P}_{Q_{(Z_i)_{i=1}^n}}(\mathcal{R}(h_n) > \gamma(n,\delta)) < \delta$. La fonction γ peut dépendre de ℓ, $\mathcal{H}$, Q_Z et de constantes universelles, et pour δ fixé, $\gamma(n,\delta)$ décroît avec n. Parfois les résultats sont plus faibles au sens où ils peuvent ne pas être vrais pour tout $\delta \in (0,1)$ et peuvent nécessiter que n soit plus grand qu'un certain seuil. Même si la question de la prédiction est importante, un pan de la recherche en apprentissage statistique s'intéresse également aux qualités de h_n en termes d'estimation. Le critère retenu pour évaluer la qualité de l'estimation est la distance $\|h_n - h^*\|_H$ pour une norme qui diffère en général de celle utilisée pour la régularisation (de nombreux exemples de $\|\cdot\|_H$ et $\|\cdot\|_R$ sont donnés dans [5]). Pour établir un résultat d'estimation, notons que h^* doit être unique ou il doit *a minima* être possible de choisir de manière unique un des minimiseurs de (1.3). Au cours des 20 dernières années, plusieurs conditions ont été proposées pour relier $\|h_n - h^*\|_H$ et $\mathcal{R}(h_n)$ et ainsi directement obtenir une inégalité oracle d'estimation à partir de celle de prédiction ([106, 134, 5, 45])

$$\mathbb{P}_{Q_{(Z_i)_{i=1}^n}}(||h_n - h^*||_H > \gamma(n,\delta)) < \delta, \quad \forall(\delta, n) \in (0,1)\times\mathbb{N}^*. \tag{1.5}$$

L'interprétation économétrique de la MRER et le besoin d'outils supplémentaires pour traiter la question de l'endogénéité

De nombreux modèles économétriques peuvent s'écrire en utilisant le cadre de la MRER présentée dans le paragraphe précédent mais les raisons pour utiliser la MRER en économétrie se distinguent de celles mises en avant en apprentissage statistique. Nous illustrons cela en nous intéressant aux modèles de régression à la moyenne et à la médiane, *i.e* nous supposons que les données sont engendrées selon l'équation $Z_o = h^{**}(Z_e) + \epsilon$ soit sous la contrainte $\mathbb{E}_{Q_{\epsilon|Z_e}}[\epsilon \mid Z_e] = 0$, soit sous la contrainte $med(Q_{\epsilon|Z_e}) = 0$. Sous la première contrainte, h^{**} est la vraie espérance conditionnelle de Z_o sachant Z_e et elle vérifie $h^{**} = \operatorname{argmin}_{h:\mathbb{E}_{Q_{Z_e}}[h(Z_e)^2]<+\infty} \mathbb{E}_{Q_Z}[(Z_o - h(Z_e))^2]$. Sous la seconde contrainte, h^{**} est la vraie médiane conditionnelle de Z_o sachant Z_e et elle satisfait $h^{**} = \operatorname{argmin}_{h:\mathbb{E}_{Q_{Z_e}}[|h(Z_e)|]<+\infty} \mathbb{E}_{Q_Z}[|Z_o - h(Z_e)|]$. La théorie économique fournit souvent des contraintes naturelles sur h^{**} telles que la monotonie, la convexité/concavité ou la régularité. Définissons $\mathcal{C}$ l'ensemble de toutes les fonctions mesurables de $\mathcal{Z}_e$ vers $\mathcal{Z}_o$ qui satisfont des contraintes dictées par la théorie économique. La classe de fonctions peut alors être choisie égale à $\mathcal{H} = \left\{h : \mathbb{E}_{Q_{Z_e}}[h(Z_e)^2] < +\infty\right\} \cap \mathcal{C}$ dans le cas de la régression à la moyenne ou $\mathcal{H} = \left\{h : \mathbb{E}_{Q_{Z_e}}[|h(Z_e)|] < +\infty\right\} \cap \mathcal{C}$ dans le cas de la régression à la médiane. La régression à la moyenne s'inscrit dans le cadre de la MRER en choisissant $\ell(Z_o, Z_e, h) = (Z_o - h(Z_e))^2$. Il en va de même de la régression à la médiane en choisissant $\ell(Z_o, Z_e, h) = |Z_o - h(Z_e)|$. Ce qui distingue l'économétrie de l'apprentissage statistique est le fait que ℓ est imposée par le paramètre d'intérêt en économétrie et n'est donc pas choisie: si nous nous intéressons à la fonction de régression à la moyenne, ℓ est nécessairement la perte des moindres carrés. De plus, $\mathcal{H}$ est choisie pour refléter des contraintes justifiées d'un point de vue économique plutôt que pour des raisons computationnelles. Pour rendre le lien entre économétrie et apprentissage statistique encore plus clair, il est utile de remarquer que (1.3) et (1.4) sont formellement équivalents à la classe des M-estimateurs régularisés, un nom qui est vraisemblablement plus familier en économétrie.

La MRER est un cadre très général qui n'est toutefois pas très adapté pour traiter d'une question fondamentale en économétrie: l'endogénéité. Ce concept saisit l'idée que certaines variables qui influencent à la fois la variable expliquée Z_o et les variables explicatives observées Z_e peuvent ne pas être observables par l'économètre. Dans ce cas, utiliser seulement Z_e pour expliquer Z_o ne permet pas *a priori* d'identifier et donc d'estimer les paramètres d'intérêt du modèle. Pour outrepasser cette difficulté, une approche standard revient à trouver des variables additionnelles appelées instruments qui ont un impact sur Z_o seulement à travers Z_e. Nous ne détaillons pas ici les raisons formelles derrière le manque de compatibilité entre l'endogénéité et la M-estimation. Intuitivement, nous pouvons quand même dire que la M-estimation est basée sur un argument de projection qui ne se lie pas bien aux techniques permettant de corriger l'endogénéité (en dehors des modèles linéaires tout du moins; voir [62] pour plus d'éléments). En présence d'endogénéité, il est en fait plus naturel de caractériser les paramètres d'intérêt en cherchant le zéro d'un ensemble judicieux de conditions de moments ([37]). Pour le voir, attardons-nous sur le modèle de régression à la médiane. Nous supposons désormais que le modèle prend la forme $Z_o = h^{**}(Z_e) + \epsilon$ avec $med(Q_{\epsilon|Z_e}) \neq 0$ mais $med(Q_{\epsilon|X}) = 0$. Dans ce modèle, au moins une composante de Z_e est liée à ϵ ce qui explique pourquoi la restriction $med(Q_{\epsilon|Z_e}) = 0$ ne tient plus. Le vecteur X contient tous les instruments plus les éléments de Z_e qui ne violent pas l'hypothèse initiale sur la médiane de $Q_{\epsilon|Z_e}$. Il est possible de montrer que le modèle de régression à la moyenne peut s'exprimer comme

$$\mathbb{E}_{Q_{Z|X}}\left[\mathbb{1}\left\{Z_o \leq h(Z_e)\right\} \mid X\right] = 0 \;\; Q_X - a.s \iff h = h^{**}.$$

Cet example justifie de s'intéresser à une classe de modèles alternative à la M-estimation

$$\mathbb{E}_{Q_{Z|X}}\left[\rho(Z,h) \mid X\right] = 0 \;\; Q_X - a.s \iff h = h^{**}, \tag{1.6}$$

où ρ est un vecteur fini-dimensionnel de fonctions connues. La relation (1.6) peut encore se réécrire $h^{**} = \operatorname{argmin}_{h\in\mathcal{H}} \mathbb{E}_{Q_X}\left[\|\mathbb{E}_{Q_{Z|X}}\left[\rho(Z,h) \mid X\right]\|^2\right]$, avec $\|\cdot\|$ la norme euclidienne. Si des contraintes naturelles peuvent être imposées sur h^{**}, il est utile de doter $\mathcal{H}$ d'une norme de régularisation $\|\cdot\|_R$ qui rend ces contraintes saillantes. La contrepartie empirique du problème devient

$$h_n \in \operatorname*{argmin}_{h\in\mathcal{H}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \|\mathbb{E}_n\left[\rho(Z,h) \mid X = X_i\right]\|^2 + \alpha\|h\|_R^p \right\}. \tag{1.7}$$

La quantité $\mathbb{E}_n\left[\rho(Z,h) \mid X = \cdot\right]$ correspond à un estimateur de la fonction $\mathbb{E}_{Q_{Z|X}}\left[\rho(Z,h) \mid X = \cdot\right]$. La procédure d'estimation (1.7) est appelée de manière peu élégante la méthode des moments généralisée régularisée (MMGR). Quand la classe de fonctions $\mathcal{H}$ est paramétrique, il est en général inutile d'ajouter un terme de régularisation. Quand $\mathcal{H}$ est paramétrique, h^{**} peut même être identifié à l'aide d'un nombre fini de moments inconditionnels dans certains cas, *i.e* $\mathbb{E}_{Q_Z}\left[\rho(Z,h)\right] = 0 \iff h = h^{**}$. Dans ce cadre particulier, il existe une connexion naturelle entre la M-estimation et la méthode des moments généralisée. Pour expliquer cette connexion, nous supposons pour simplifier que $h(Z_e) = Z_e'\beta$. Sous certaines conditions sur $\mathbb{E}_{Q_Z}\left[\ell(Z_o, Z_e, \langle\cdot,\beta\rangle)\right]$ incluant la différentiabilité en β, (1.3) est équivalent à

$$\frac{\partial}{\partial\beta}\mathbb{E}_{Q_Z}\left[\ell(Z_o, Z_e, \langle\cdot,\beta\rangle)\right] = 0 \iff \beta = \beta^{**}.$$

S'il existe une fonction $\rho : (z_o, z_e, \beta) \mapsto \rho(z, \langle\cdot,\beta\rangle)$ telle que pour tout β $\frac{\partial}{\partial\beta}\mathbb{E}_{Q_Z}\left[\ell(Z_o, Z_e, \langle\cdot,\beta\rangle)\right] = \mathbb{E}_{Q_Z}\left[\rho(Z, \langle\cdot,\beta\rangle)\right]$, alors le problème de M-estimation a été traduit dans le cadre de la méthode des moments généralisée. Les estimateurs de la méthode des moments généralisée qui sont obtenus à partir d'un M-estimateur sont appelés Z-estimateurs (ceci est expliqué dans le chapitre 5 de [136]).

La procédure (1.7) a été étudiée dans de nombreuses contributions dans la littérature économétrique, [37] étant une pierre angulaire. Contrairement à l'approche de l'apprentissage statistique, les propriétés de prédiction de h_n ne sont pas d'une importance centrale. La plupart des résultats reviennent à prouver que $\|h_n - h^{**}\|_H$ converge vers zéro à une vitesse suffisamment rapide, pour une norme $\|\cdot\|_H$ possiblement différente de $\|\cdot\|_R$. Les résultats sont asymptotiques la plupart du temps au sens où il existe peu d'articles proposant des inégalités oracles d'estimation.

La construction d'ensembles de confiance en économétrie et en statistique nonparamétrique

En principe, des ensembles de confiance peuvent être construits sans se baser sur un estimateur du paramètre d'intérêt. En pratique cependant, les ECs sont pratiquement toujours construits à partir d'un estimateur et il y a *de facto* un lien fort entre l'estimation et la construction d'ECs. Pour comprendre ce lien, nous regardons deux cas: la construction d'un EC pour h^{**} dans le modèle (2.3) (avec $h^* = h^{**}$) et la construction d'un EC pour une fonctionnelle de h^{**} notée $\varphi(h^{**})$.

Le premier exemple est typiquement ce qui intéresse les chercheurs en statistique nonparamétrique ([79]). Si une inégalité oracle du même type que (1.5) existe et $\gamma(\cdot,\cdot)$ ne dépend pas de h^{**}, un EC de niveau $1-\delta$ valide pour chaque n uniformément sur $\mathcal{H}$ peut être construit en identifiant tous les $h \in \mathcal{H}$ tels que $\|h_n - h\|_H \leq \gamma(n,\delta)$. C'est un résultat théorique attrayant car il s'applique à une large classe de problèmes statistiques mais son implémentation directe est souvent difficile: la norme $\|\cdot\|_H$ peut être pénible à calculer et la recherche de tous les hs qui appartiennent à l'EC peut être computationnellement très lourde; la fonction $\gamma(\cdot,\cdot)$ dépend en général de constantes qui sont soit inconnues, soit très grandes ou doivent être estimées; quand bien même les deux difficultés précédentes n'apparaissent pas, les ECs

fondés sur des inégalités oracles peuvent avoir un diamètre qui est trop large asymptotiquement dans un sens que nous explicitons plus bas. Pour rendre ces questions plus parlantes, nous nous ramenons maintenant à un problème très simple. Nous souhaitons construire un intervalle de confiance pour $\mathbb{E}_{Q_{Z_o}}[Z_o]$ en utilisant n tirages *i.i.d* de loi Q_{Z_o}. Nous supposons que la variance de Q_{Z_o} est finie et connue, fixée égale à V. Une application de l'inégalité de Bienaymé-Chebyshev (voir chapitre 2 dans [136]) permet d'obtenir l'inégalité oracle suivante: pour tout $n \geq 1$ et $\delta \in (0,1)$, $\mathbb{P}_{Q_{Z_o}^{\otimes n}}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_{o,i} - \mathbb{E}_{Q_{Z_o}}[Z_o]\right| > \sqrt{\frac{V}{n\delta}}\right) < \delta$. L'intervalle $I_n^{\delta,1} := \left[\frac{1}{n}\sum_{i=1}^n Z_{o,i} - \sqrt{V/n\delta}, \frac{1}{n}\sum_{i=1}^n Z_{o,i} + \sqrt{V/n\delta}\right]$ satisfait pour chaque $n \geq 1$

$$\inf_{\mathbb{E}_{Q_{Z_o}} \in \mathbb{R}} \mathbb{P}_{Q_{Z_o}^{\otimes n}}\left(I_n^{\delta,1} \ni \mathbb{E}_{Q_{Z_o}}[Z_o]\right) \geq 1 - \delta.$$

L'hypothèse que la variance est connue égale à V est malheureusement trop restrictive en pratique. Une solution serait de: i) supposer que la vraie variance est inconnue mais bornée supérieurement par V connu ce qui ne changerait pas le résultat, ii) remplacer V par la variance empirique mais la validité nonasymptotique de $I_n^{\delta,1}$ ne tiendrait plus. Notons que i) n'est intéressant en pratique que s'il existe une borne naturelle et assez petite sur la variance. Nous insistons cependant sur le fait qu'une borne supérieure sur la variance (ou sur des moments plus élevés) est nécessaire pour construire des intervalles de confiance valides de manière strictement nonasymptotique.

Le second exemple est central en économétrie où le paramètre d'intérêt est souvent non pas h^{**} mais une fonctionnelle de ce dernier (voir l'introduction du chapitre 3 pour de nombreuses références et des exemples de fonctionnelles intéressantes en économie). En économétrie, l'approche pour construire des intervalles de confiance est principalement asymptotique: la méthode usuelle consiste à trouver une suite (aléatoire) r_n telle que la loi de $r_n(\varphi(h_n) - \varphi(h^{**}))$ converge vers une loi $\mathcal{N}(0,1)$. Soit $q_{\mathcal{N}(0,1)}(1-\delta/2)$ le quantile $1-\delta/2$ de la loi $\mathcal{N}(0,1)$. Nous pouvons montrer que l'intervalle $I_n^{\delta,2} := \left[\varphi(h_n) - q_{\mathcal{N}(0,1)}(1-\delta/2)/r_n, \varphi(h_n) + q_{\mathcal{N}(0,1)}(1-\delta/2)/r_n\right]$ est asymptotiquement de niveau $1-\delta$ ponctuellement sur Θ. $I_n^{\delta,2}$ satisfait les critères d'optimalité présentés plus haut, en particulier la probabilité que $\varphi(h^{**})$ appartienne à $I_n^{\delta,2}$ tend vers $1-\delta$ pour tout $\delta \in (0,1)$ et $h^{**} \in \mathcal{H}$. Le principal défaut de $I_n^{\delta,2}$ est que son comportement est incontrôlé pour tout n fini et qu'il n'est pas honnête au sens donné plus haut sans restriction supplémentaire (voir [97]).

Revenons au premier exemple. En utilisant les mêmes arguments que pour $I_n^{\delta,2}$, nous pouvons construire un intervalle de confiance ponctuellement valide asymptotiquement de la manière suivante $I_n^{\delta,3} := \left[\frac{1}{n}\sum_{i=1}^n Z_{o,i} - q_{\mathcal{N}(0,1)}(1-\delta/2)\sqrt{V/n}, \frac{1}{n}\sum_{i=1}^n Z_{o,i} + q_{\mathcal{N}(0,1)}(1-\delta/2)\sqrt{V/n}\right]$. Si nous calculons le ratio des longueurs de $I_n^{\delta,1}$ et $I_n^{\delta,3}$ et étudions sa limite en probabilité, nous remarquons que $diam(I_n^{\delta,1})/diam(I_n^{\delta,3}) \to 1/(q_{\mathcal{N}(0,1)}(1-\delta/2)\delta)$. Nous pouvons montrer que cette limite est plus grande que 1 pour tout $\delta \in (0, 1/2)$, ce qui implique que $I_n^{\delta,1}$ est de niveau asymptotique strictement plus grand que $1-\delta$ et est donc conservateur.

La discussion précédente souligne le fait qu'il est difficile de combiner optimalité asymptotique et honnêteté. Ces deux notions ne sont toutefois pas incompatibles et une littérature traitant de cette question a éclos ([87, 123, 122]). Dans les années récentes, plusieurs économètres ont été prolifiques dans ce champ de recherche et ont proposé des méthodes intéressantes tant sur le plan théorique que pratique.

Relâcher l'hypothèse *i.i.d* a de l'importance en économétrie

Il y a de nombreuses raisons naturelles d'aller au-delà de l'hypothèse *i.i.d*. La dimension temporelle d'un problème est vraisemblablement la première raison: quand le temps joue un rôle dans l'analyse, ce qui est le cas avec les données de panel, il est très plausible que les données soient dépendantes au cours du temps (du fait de phénomènes de persistance) et que la loi des observations se modifie à

plus ou moins long terme. De ce fait, les observations ne sont plus ni indépendantes ni identiquement distribuées. Nous ne traitons pas plus avant la question du temps dans la modélisation statistique car dans les chapitres qui suivent notre attention se porte sur des modèles où le temps n'est pas un élément clé.

Même dans le cas de données en coupe (*i.e* des données qui ne sont pas indicées par le temps), l'hypothèse *i.i.d* est souvent considérée peu crédible en pratique. Prenons un exemple simple: nous observons un échantillon de n travailleurs et nous disposons d'informations sur leur zone d'emploi et leur secteur d'activité. En économétrie appliquée, il est courant d'autoriser des chocs agrégés au niveau de la zone géographique et du secteur d'activité ([1, 27, 110]). Le but est de construire des ECs qui sont *robustes* à la présence de tels chocs. Des ECs sont dits robustes s'ils ont le bon niveau (asymptotique) que les données soient *i.i.d* ou pas. L'hypothèse *i.i.d* est également peu crédible avec des données d'interaction, c'est-à-dire des données qui proviennent des interactions entre les individus d'une même population. Dans ce cadre, un jeu de données prend typiquement la forme d'une suite doublement indicée $(W_{i,j})_{1\leq i\neq j\leq n}$ où $W_{i,j}$ est l'observation relative à la paire formée par les individus i et j. Ces notions de dépendance en coupe existent dans d'autres domaines comme la statistique spatiale ou l'analyse des réseaux. Néanmoins, dans ces deux derniers champs, la dépendance est le principal sujet d'intérêt, ce qui signifie qu'un modèle est stipulé quant à la structure de dépendance et le but est d'estimer les paramètres dudit modèle. En économétrie (tout du moins pour les questions qui nous intéressent) le but est assez différent: la dépendance est principalement vue comme un terme de nuisance dont il doit être tenu compte pour faire de l'inférence de manière valide sur d'autres paramètres. La dépendance en coupe est au coeur du chapitre 4.

Dans le paragraphe précédent, nous n'avons pas relâché l'hypothèse que les observations sont identiquement distribuées. Nous ne levons jamais cette contrainte dans les chapitres qui suivent et nous la considérons même comme assez fondamentale (à l'exception du cas des données indexées par le temps): il semble en effet assez naturel de supposer que deux individus issus d'un même échantillon - aussi différents soient-ils en termes de niveau d'éducation et de salaire par exemple - sont simplement deux réalisations distinctes issues d'une même loi. Certains économètres et statisticiens ont une approche différente: ils prennent les variables explicatives observées $(Z_{e,i})_{i=1}^n$ comme déterministes ce qui conduit à considérer un échantillon non identiquement distribué (voir le chapitre 2.8 dans [136]).

Causalité et machine learning

La causalité est un des piliers de la discipline économétrique. Cette notion a été popularisée en économétrie à la suite d'un article de Donald Rubin ([124]). Elle repose sur une expérience de pensée: il existe deux états de la nature (notés 0 et 1) et chaque individu est placé dans un de ces états. Les individus se voient attribuer une variable expliquée $Z_o(0)$ ou $Z_o(1)$ selon l'état dans lequel ils se trouvent. Au niveau individuel, l'effet causal du passage d'un état à un autre est simplement la différence $Z_o(1) - Z_o(0)$. Pourquoi est-ce que la causalité est intéressante en économétrie? C'est un cadre pratique pour modéliser l'impact d'une politique publique au niveau agrégé. Si le gouvernement pouvait observer $Z_o(1) - Z_o(0)$ pour tout le monde, ce gouvernement pourrait mesurer l'effet de faire changer les individus d'état selon une règle donnée. Dans ce contexte, mettre en place une politique publique est équivalent à l'action de faire changer les individus d'état.

En réalité, le gouvernement observe soit $Z_o(1)$ soit $Z_o(0)$ mais jamais les deux: le cadre causal est un exemple d'un problème statistique dit de données manquantes ([121]). En notant D l'état dans lequel se trouve un individu, le gouvernement observe seulement $Z_o = DZ_o(1) + (1 - D)Z_o(0)$. Sans restriction supplémentaire, il est seulement possible d'identifier $Q_{Z_o(1)|D=1}$ et $Q_{Z_o(0)|D=0}$. La restriction

supplémentaire $(Z_o(1), Z_o(0)) \perp\!\!\!\perp D$ nous assure que $Q_{Z_o(1)|D=1} = Q_{Z_o(1)}$ et $Q_{Z_o(0)|D=0} = Q_{Z_o(0)}$. Nous nous référons à [81] pour une présentation détaillée de la question de l'identification dans le cadre causal de Rubin. L'identification de $Q_{Z_o(0)}$ et $Q_{Z_o(1)}$ permet de calculer l'impact moyen associé au traitement D: $\mathbb{E}_{Q_{Z_o(0),Z_o(1)}}[Z_o(1) - Z_o(0)]$, ou le changement au niveau du δ-ème quantile: $q_{Q_{Z_o(1)}}(\delta) - q_{Q_{Z_o(0)}}(\delta)$. A l'inverse, l'identification de $Q_{Z_o(0)}$ et $Q_{Z_o(1)}$ n'est pas suffisante pour obtenir le δ-ème quantile de l'effet de traitement $q_{Q_{Z_o(1)-Z_o(0)}}(\delta)$. Pour avoir l'égalité $q_{Q_{Z_o(1)}}(\delta) - q_{Q_{Z_o(0)}}(\delta) = q_{Q_{Z_o(1)-Z_o(0)}}(\delta)$, nous devons imposer que le rang d'un individu sous la loi $Q_{Z_o(0)}$ est le même que sous la loi $Q_{Z_o(1)}$ (c'est la propriété d'invariance des rangs, *cf* [65]).

Dans le reste de ce paragraphe, nous nous concentrons sur le paramètre $\mathbb{E}_{Q_{Z_o(0),Z_o(1)}}[Z_o(1) - Z_o(0)]$ que nous notons τ. Un des désavantages de l'hypothèse $(Z_o(1), Z_o(0)) \perp\!\!\!\perp D$ est sa non-testabilité. Elle est souvent remplacée par $(Z_o(1), Z_o(0)) \perp\!\!\!\perp D \mid Z_e$ qui n'est pas testable non plus mais strictement plus faible. Sous cette dernière condition, il est possible de montrer ([81]) que $\tau = \mathbb{E}_{Q_{Z_e}}\left[\mathbb{E}[Z_o \mid D = 1, Z_e] - \mathbb{E}[Z_o \mid D = 0, Z_e]\right]$. Le terme de droite dépend seulement de variables observées. Les deux tâches qui intéressent principalement un économètre sont: i) l'estimation de et l'inférence sur τ, ii) tester l'hétérogénéité des effets de traitement pour différents profils individuels z_e. Ce deuxième objectif revient à tester si $\mathbb{E}_{Q_{(Z_o(0),Z_o(1))|Z_e}}[Z_o(1) - Z_o(0) \mid Z_e = z_1] = \mathbb{E}_{Q_{(Z_o(0),Z_o(1))|Z_e}}[Z_o(1) - Z_o(0) \mid Z_e = z_2]$ quand $z_1 \neq z_2$. Pour chacune des deux tâches précédentes, il faut estimer dans un premier temps les fonctions $\mathbb{E}[Z_o \mid D = 1, Z_e = \cdot]$ et $\mathbb{E}[Z_o \mid D = 0, Z_e = \cdot]$ (évaluées seulement aux points z_1 et z_2 pour le deuxième objectif). Comment estimer ces fonctions de manière flexible? Une possibilité est d'avoir recours aux outils classiques de statistique nonparamétrique tels que la régression de Nadaraya-Watson ou la régression linéaire locale ([135]). Les garanties théoriques de ces méthodes ont été établies depuis plusieurs décennies ([59, 70]). Leur principale limite est leur mauvaise performance en pratique quand la dimension de Z_e est grande. En revanche, les techniques plus récentes issues du machine learning, telles les forêts aléatoires ou les réseaux de neurones profonds, sont très performantes sur simulations et en pratique quand la dimension de Z_e est grande, mais leurs propriétés théoriques sont bien moins connues. Des efforts récents de recherche tant en économétrie qu'en apprentissage statistique ont permis des avancées théoriques sur les algorithmes de machine learning: le théorème 3 dans [71] montre la normalité asymptotique d'un estimateur de τ basé sur un réseau de neurones profond, [138] prouve la normalité asymptotique d'une méthode utilisant les forêts aléatoires pour estimer $\mathbb{E}_{Q_{(Z_o(0),Z_o(1))|Z_e}}[Z_o(1) - Z_o(0) \mid Z_e = z_e]$ pour un z_e fixé. Il est intéressant de constater que les propriétés théoriques ne sont pas très différentes de celles d'outils de statistique nonparamétrique plus anciens: les résultats actuels pour les réseaux de neurones profonds sont valides pour les mêmes classes de fonctions que pour des outils plus classiques et la performance théorique de ces réseaux est elle-aussi sensiblement impactée par la dimension de Z_e; les forêts aléatoires peuvent approcher des fonctions qui sont moins régulières mais leur performance théorique se dégrade malgré tout avec la dimension de Z_e.

Résumé du chapitre 3

Dans ce chapitre, nous nous concentrons sur le problème générique donné par (1.6). Comme expliqué plus haut, de nombreux articles de recherche (et même la majorité) qui traitent de ce problème proposent des procédures d'estimation basées sur (1.7) ([3], [112], [20], and [37] pour n'en citer que quelques-uns). Il existe d'autres façons de construire des estimateurs pour cette classe de problèmes et nous nous intéressons à la famille d'estimateurs dits de vraisemblance empirique généralisée que nous appelons GEL par la suite ([113], [99]). Pour présenter les estimateurs GEL, il est commode de partir d'une version simplifiée de (1.6): nous supposons que h est remplacée par un paramètre fini-dimensionnel $\beta \in \mathcal{B}$ et la

vraie valeur du paramètre β^{**} est telle que $\mathbb{E}_{Q_Z}[\rho(Z,\beta)] = 0 \iff \beta = \beta^{**}$. [113, 99] expliquent que β^{**} est également identifié par

$$\beta^{**} = \underset{\beta\in\mathcal{B}}{\operatorname{argmin}} \sup_{\lambda\in\Lambda(\beta,Q_Z)} \mathbb{E}_{Q_Z}\left[\psi_\gamma(\lambda'\rho(Z,\beta))\right], \tag{1.8}$$

avec $\Lambda(\beta, Q_Z) := \bigcap_{z\in supp(Q_Z)} \{\lambda : \psi_\gamma(\lambda'\rho(z,\beta))$ existe$\}$ et $\psi_\gamma : u \mapsto \frac{2}{\gamma}\left[-(\gamma+1)\frac{u+1}{2}\right]^{\frac{\gamma}{\gamma+1}} - \frac{2}{\gamma(\gamma+1)}$. En prenant la contrepartie empirique du problème de point-selle précédent, nous obtenons un estimateur pour chaque fonction ψ_γ. Nous pouvons définir ainsi la famille des estimateurs GEL. Les membres les plus connus de cette famille sont: l'estimateur associé à l'*Empirical Likelihood* (EL) qui fut popularisée par [117], l'*Exponential Tilting* ([100]) et l'estimateur dit *continuously updating* (CUE) de [88]. Les idées ci-dessus s'appliquent aux problèmes de la forme (1.6). [93] montre que (1.6) peut être reformulé sous la forme (1.8) avec un nombre d'égalités de moment qui diverge avec n: h^{**} est l'unique valeur du paramètre qui satisfait pour tout $n \geq 1$

$$h^{**} = \underset{h\in\mathcal{H}}{\operatorname{argmin}} \sup_{\lambda\in\Lambda(h,Q_{Z,X})} \mathbb{E}_{Q_{Z,X}}\left[\psi_\gamma(\lambda'\rho(Z,h)) \otimes q_{K_n}(X)\right], \tag{1.9}$$

avec $\otimes$ le produit de Kroneker et $q_{K_n}(\cdot)$ un vecteur de dimension croissante K_n composé de fonctions bien choisies. [101, 99] proposent une adaptation plus directe: ils montrent que h^{**} vérifie

$$h^{**} = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\, \mathbb{E}_{Q_Z}\left[\sup_{\lambda\in\Lambda(h,Q_{Z|X})} \mathbb{E}_{Q_{Z|X}}\left[\psi_\gamma(\lambda'\rho(Z,h)) \mid X\right]\right], \tag{1.10}$$

où $\Lambda(h, Q_{Z|X=x}) := \bigcap_{z\in supp(Q_{Z|X=x})} \{\lambda : \psi_\gamma(\lambda'\rho(z,h))$ existe$\}$. Remarquons que même lorsque h se réduit à un paramètre fini-dimensionnel (comme dans [101, 99]), $\mathbb{E}_{Q_{Z|X}}[\cdot \mid X = \cdot]$ est nonparamétrique sans plus de restriction. Pour construire des estimateurs GEL, les articles cités plus haut se basent sur la contrepartie empirique de (1.9) ou (1.10) et utilisent un estimateur nonparamétrique pour approximer $\mathbb{E}_{Q_{Z|X}}[\cdot \mid X = \cdot]$.

Il existe très peu de contributions où h est autorisée à être de dimension infinie, les deux principales étant [116] et [40]. La classe de fonctions $\mathcal{H}$ est toujours choisie comme un sous-ensemble d'un espace métrique doté d'une norme $\|\cdot\|_H$. Cet espace métrique est le plus souvent l'espace des fonctions de carré intégrable par rapport à la mesure de Lebesgue ou l'espace des fonctions uniformément bornées par rapport à la même mesure. Dans [116], l'auteur se concentre sur des modèles où ρ dépend de manière régulière de h et il étudie le comportement de l'estimateur EL construit à partir de (1.10). Il utilise une méthode de Nadaraya-Watson pour estimer $\mathbb{E}_{Q_{Z|X}}[\cdot \mid X = \cdot]$. Ses principaux résultats sont la convergence de son estimateur en norme $\|\cdot\|_H$ et la normalité asymptotique d'une certaine fonctionnelle de son estimateur. Des restrictions assez fortes sont imposées sur la classe $\mathcal{H}$ pour contourner le besoin de régulariser la procédure d'estimation ce qui est une limite de cet article. Dans [40], les auteurs étudient le comportement de l'ensemble des estimateurs GEL construits à partir de (1.9) pour un modèle particulier, à savoir la régression quantile instrumentale nonparamétrique (NPQIV) ([41]). Dans le NPQIV, ρ ne dépend pas de h de façon régulière et n'est donc pas traité dans [116]. Dans [40], des classes de fonctions $\mathcal{H}$ plus larges que dans [116] sont par ailleurs considérées grâce à un terme de régularisation qui est introduit dans la procédure d'estimation. Les principaux résultats dans [40] sont la convergence des estimateurs GEL en norme $\|\cdot\|_H$ avec une vitesse explicite ainsi que la normalité asymptotique d'une large classe de fonctionnelles de ces estimateurs.

Dans le chapitre 3, nous étudions les propriétés de l'ensemble des estimateurs GEL pour une classe de fonctions ρ, et donc de modèles, qui englobe ceux couverts par [116] et [40]. Tout comme [40], nous proposons une procédure d'estimation régularisée et considérons des classes de fonctions $\mathcal{H}$

plus générales que [116]. Notre approche se distingue de celles de [116] et [40] car nous utilisons une version un peu modifiée de (1.10) pour construire nos estimateurs. Comme souligné précédemment, le recours à la régularisation ne signifie pas que $\mathcal{H}$ peut être choisie arbitrairement grande: nous supposons que $\mathcal{H}$ contient des fonctions de carré intégrable, différentiables jusqu'à un certain ordre avec des dérivées partielles de carré intégrable elles aussi. Soit $\|\cdot\|_{L_2(leb)}$ la norme associée à l'ensemble des fonctions de carré intégrable par rapport à la mesure de Lebesgue. Dans notre travail, nous prouvons la convergence des estimateurs GEL en norme $\|\cdot\|_{L_2(leb)}$ et nous établissons une borne supérieure sur la vitesse à laquelle $\mathbb{E}_{Q_X}\left[\|\mathbb{E}_{Q_{Z|X}}\left[\rho(Z,h_n)\right]\|^2\right]$ converge vers 0. Nous obtenons une vitesse lente de convergence qui requiert l'existence d'un nombre limité de moments de ρ et nous montrons que la vitesse peut être améliorée à condition que les moments de ρ existent jusqu'à un ordre plus élevé. Nous expliquons comment ces résultats peuvent être utilisés pour obtenir la vitesse de convergence de nos estimateurs en norme $\|\cdot\|_{L_2(leb)}$. Dans le chapitre 3, nous rappelons notamment que pour arriver à ce dernier résultat, la clé est de contrôler le ratio $\|h-h^{**}\|_{L_2(leb)}/\mathbb{E}_{Q_X}\left[\|\mathbb{E}_{Q_{Z|X}}\left[\rho(Z,h)\right]\|^2\right]$ uniformément en h dans un voisinage bien choisi de h^{**}. Ce ratio mesure l'écart entre une norme au numérateur et ce qui s'apparente à une norme plus faible au dénominateur. Le supremum de ce ratio caractérise à quel point le problème est mal-conditionné. Nous parlons d'un problème *ill-posed* en anglais ([37]). Une littérature très riche et encore active a proposé des conditions suffisantes pour contrôler le degré de *ill-posedness* du modèle statistique d'intérêt (voir [39, 33] qui dressent des revues de littérature très complètes). Comme expliqué à la fin du chapitre 3, nous pensons qu'il y a encore matière à améliorer les conditions existantes proposées pour contrôler le degré de *ill-posedness*. Ceci est clairement un axe de recherche futur que nous souhaitons explorer et dont l'intérêt dépasse le cadre des modèles statistiques vérifiant (2.6).Nous pouvons citer d'autres extensions possibles de nos résultats actuels: i) montrer la normalité asymptotique pour la même classe de fonctionnelles que [40]; ii) dans un esprit plus purement statistique, construire des inégalités oracles sur la performance de nos estimateurs.

Résumé du chapitre 4

Même dans le cas de données en coupe, l'hypothèse *i.i.d* peut être trop restrictive. En pratique, il est souvent plausible que les données soient affectées par plusieurs chocs agrégés inobservés: supposons que nous observions plusieurs variables au niveau secteur d'activité-zone géographique. Les données peuvent s'écrire $(Z_{i_1,i_2})_{1\leq i_1\leq n_1,1\leq i_2\leq n_2}$, avec n_1 (*resp.* n_2) le nombre de secteurs (*resp.* zones géographiques). Les observations correspondent à des cellules secteur-zone géographique et elles sont *a priori* liées entre elles dès qu'elles partagent le même secteur ou la même zone géographique du fait de chocs économiques potentiellement inobservés à ces niveaux. Nous parlons en général de données avec une structure en grappe dans les dimensions secteur et zone. C'est donc un exemple de ce qui s'appelle la dépendance multiple en grappe. Les données polyadiques sont un autre type de données qui présente naturellement une structure de dépendance: ces données proviennent des interactions entre les individus d'une même population les uns avec les autres. Les données sur les relations entre des paires d'individus sont appelées dyadiques et sont les plus courantes. Les données dyadiques peuvent être représentées sous la forme $(Z_{i_1,i_2})_{1\leq i_1\neq i_2\leq n}$. Intuitivement, la dépendence polyadique devrait être plus forte que celle en grappe multiple: dans le premier cas, les observations sont liées du fait de chocs issus d'une unique population alors que dans le deuxième cas, les chocs proviennent de deux sources distinctes. Pour modéliser ces idées, nous faisons l'hypothèse que les données sont jointement échangeables dans le cas polyadique et séparablement échangeables dans le cas de dépendance multiple en grappe. Ces deux notions d'échangeabilité sont présentées de manière détaillée dans [96]. Ces hypothèses sont puissantes car elles permettent d'utiliser des résultats probabilistes très profonds et utiles ([89, 4, 95]) qui

assurent que les données puissent être représentées en fonction d'un ensemble de chocs inobservés indépendants dans les différentes dimensions de dépendance. Bien que l'échangeabilité séparable soit un sous-cas de l'échangeabilité jointe, nous devons quand même traiter la dépendance multiple en grappe à part: le nombre différent de grappes dans chaque dimension rend le problème plus compliqué. Il est à noter que l'échangeabilité implique que les données demeurent identiquement distribuées: la dépendance que nous introduisons est donc très différente de celle qui surgit dans les séries temporelles.

Quand l'hypothèse d'échangeabilité remplace celle que les données sont *i.i.d*, la construction des estimateurs n'est pas affectée. Cependant, il faut quand même modifier les arguments employés pour prouver la convergence et la normalité asymptotique des estimateurs. Les résultats existants concernent essentiellement les moyennes empiriques et le modèle de régression linéaire: dans le cas jointement échangeable, la normalité asymptotique pour les moyennes empiriques remonte à [66] et celle des t-statistiques dans le modèle de régression linéaire est établie dans [131]; en présence de dépendance multiple en grappe, [109] étudie la limite en loi des moyennes simples quand le nombre de dimensions de dépendance n'est pas connu et il montre la validité asymptotique d'une procédure de bootstrap (nous définissons ce qu'est une procédure de bootstrap un peu plus loin dans ce paragraphe). Plusieurs articles proposent également des estimateurs de la variance asymptotique pour une grande classe de modèles sans prouver leur convergence ([69, 30]). Lorsque nous nous intéressons à des modèles autres que le modèle de régression linéaire, les résultats sur les moyennes empiriques ne sont en général pas suffisants. Dans le cas *i.i.d*, une approche qui a fait ses preuves consiste à contrôler le comportement asymptotique du processus empirique associé au modèle (voir [137] pour plus de détails et une définition d'un processus empirique). Nous étendons des résultats classiques sur les processus empiriques pour des données *i.i.d* aux cas de la dépendance multiple en grappe et des données polyadiques. Pour étendre ces résultats, nous devons adapter la définition d'un processus empirique. A titre d'exemple, en présence de données doublement dépendantes en grappes, le processus empirique associé à une classe de fonctions $\mathcal{F}$ est l'application aléatoire $\mathbb{G}_{n_1,n_2} : f \in \mathcal{F} \mapsto \frac{\sqrt{\min\{n_1,n_2\}}}{n_1 n_2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \left(f(Z_{i_1,i_2}) - \mathbb{E}_{Q_Z}[f(Z_{1,1})]\right)$. Avec des données dyadiques, le processus empirique prend la forme $\mathbb{G}_n : f \in \mathcal{F} \mapsto \frac{\sqrt{n}}{n(n-1)} \sum_{1 \leq i_1 \neq i_2 \leq n} \left(f(Z_{i_1,i_2}) - \mathbb{E}_{Q_Z}[f(Z_{1,2})]\right)$. La classe de fonctions $\mathcal{F}$ dépend du modèle d'intérêt. Par exemple, si nous étudions le modèle $\mathbb{E}_{Q_Z}[\rho(Z,h)] = 0 \iff h = h^{**}$, nous avons $\mathcal{F} := \{\rho(\cdot,h) : h \in \mathcal{H}\}$. Remarquons que pour chaque $f \in \mathcal{F}$, $\mathbb{G}_{n_1,n_2} f$ et $\mathbb{G}_n f$ sont asymptotiquement normaux grâce aux résultats présentés plus haut pour les moyennes empiriques. L'étude de la limite en loi du processus empirique est ainsi plus difficile que la simple vérification de la normalité asymptotique du processus pour un f fixé. Notre résultat principal est le suivant: nous montrons que les processus empiriques avec des données multiplement dépendantes en grappes ou polyadiques convergent en loi vers un processus gaussien sous les mêmes hypothèses que dans le cadre *i.i.d* mais la variance asymptotique est différente de celle obtenue dans ce dernier cas. Le processus gaussien a les propriétés suivantes: c'est une fonction aléatoire qui associe à chaque $f \in \mathcal{F}$ une variable normale centrée et de variance donnée par la formule de variance asymptotique. Ce résultat n'est pas directement utilisable pour faire de l'inférence sur un modèle statistique puisque la variance asymptotique est inconnue et doit être estimée. Au lieu de proposer un estimateur de variance, nous montrons la validité asymptotique de deux versions modifiées du bootstrap nonparamétrique ([67]) adaptées à nos schémas de dépendance. Nous expliquons à présent comment sont construites nos procédures de bootstrap dans les cas simples de la double dépendance en grappe et des données dyadiques. En présence de double dépendance en grappe, pour chaque dimension de dépendance j nous tirons n_j indices avec remise et la version bootstrap du processus empirique s'écrit $\mathbb{G}^*_{n_1,n_2} : f \in \mathcal{F} \mapsto \frac{\sqrt{\min\{n_1,n_2\}}}{n_1 n_2} \sum_{1 \leq i_1 \leq n_1} \sum_{1 \leq i_2 \leq n_2} (V^1_{i_1} V^2_{i_2} - 1) f(Z_{i_1,i_2})$, avec $V^1_{i_1}$ (*resp.* $V^2_{i_2}$) le nombre

de fois où l'indice i_1 (*resp.* i_2) est rééchantillonné. Avec des données dyadiques, nous tirons n indices avec remise et le processus bootstrap devient $\mathbb{G}_n^* : f \in \mathcal{F} \mapsto \frac{\sqrt{n}}{n(n-1)} \sum_{1 \leq i_1 \neq i_2 \leq n} (V_{i_1} V_{i_2} - 1) f(Z_{i_1,i_2})$. Dans le cadre *i.i.d*, le bootstrap nonparamétrique de base fonctionne car les observations utilisées pour le rééchantillonnage sont indépendantes. Pour nos procédures modifiées, nous ne pouvons pas rééchantillonner au niveau des observations car celles-ci ne sont pas du tout indépendantes. Nous trouvons donc un autre niveau auquel de l'indépendance apparaît: le niveau des individus qui génèrent les paires avec des données dyadiques et les deux dimensions de dépendance avec des données doublement dépendantes en grappes. La deuxième contribution de notre travail est la preuve de la validité asymptotique de nos deux procédures de bootstrap modifiées. Nous utilisons nos deux résultats principaux pour montrer la normalité asymptotique et la validité de l'inférence basée sur le bootstrap pour une grande classe de modèles nonlinéaires. Nous revisitons également la célèbre contribution empirique de [126]. Les auteurs estiment les déterminants des volumes d'échange entre pays à l'aide de variables explicatives aux niveaux pays et paire de pays telles que le produit intérieur brut ou la distance entre deux pays. Ils utilisent un modèle de pseudo maximum de vraisemblance de Poisson (PPML en anglais) et supposent que les données sont indépendantes entre paires de pays conditionnellement aux variables explicatives. Nous réestimons leur modèle et montrons qu'une fois la dépendance dyadique prise en compte, la longueur des intervalles de confiance et les p-valeurs des tests de significativité des coefficients du modèle augmentent sensiblement.

Résumé du chapitre 5

En économétrie, de nombreux paramètres d'intérêt peuvent s'écrire comme des fonctions de un ou plusieurs ratios d'espérances et/ou de covariances. Les coefficients dans une régression linéaire univariée avec ou sans endogénéité, les espérances conditionnelles et l'estimand des différences de différences avec un traitement endogène ([53]) en sont des exemples phares. Pour faire de l'inférence sur ces paramètres, l'approche économétrique standard repose sur la normalité asymptotique des moyennes empririques combinée à la méthode delta (voir le chapitre 3 de [136] pour une définition de la méthode delta). Dans notre travail, nous nous concentrons sur le cas simple d'un ratio d'espérances $\mathbb{E}_{Q_X}[X]/\mathbb{E}_{Q_Y}[Y]$ et regardons l'impact sur l'inférence d'avoir un dénominateur $\mathbb{E}_{Q_Y}[Y]$ "proche de zéro". Des résultats profonds ont déjà été montrés sur ce sujet: si le modèle n'impose pas que $\mathbb{E}_{Q_Y}[Y]$ soit séparé de zéro, il a été prouvé dans [63] que pour tout niveau de confiance un intervalle de confiance honnête au sens de (2.2) doit être de longueur infinie avec probabilité positive. Les théorèmes énoncés dans [63] s'appliquent même pour n'importe quel nombre fini d'observations n en supprimant la limite inférieure dans (2.2). La question qui nous intéresse est proche dans l'esprit du problème bien connu des variables instrumentales faibles: ce problème apparaît quand les instruments ont une corrélation presque nulle avec la variable endogène dans un modèle de régression linéaire avec endogénéité (voir [8] pour une revue de littérature récente). La littérature sur les variables instrumentales faibles a suggéré de construire des intervalles de confiance qui sont robustes à l'absence de corrélation entre les instruments et les variables endogènes en se basant sur les idées initiées dans [6]. Cette littérature a également étudié la limite en loi de plusieurs estimateurs pour le modèle linéaire en présence d'endogénéité lorsque la corrélation entre les variables endogènes et les instruments est autorisée à décroître vers zéro quand le nombre d'observations n augmente ([129]). Nous utilisons cette dernière approche pour définir la "proximité à zéro" de $\mathbb{E}_{Q_Y}[Y]$: nous autorisons $\mathbb{E}_{Q_Y}[Y]$ à dépendre de n et à décroître lorsque n augmente. Nous autorisons également $\mathbb{V}_{Q_Y}[Y]$, $\mathbb{E}_{Q_X}[X]$ and $\mathbb{V}_{Q_X}[X]$ à dépendre de n et à potentiellement décroître vers zéro. Dans ce cadre, nous établissons le comportement asymptotique de la loi de $\overline{X}_n/\overline{Y}_n - \mathbb{E}_{Q_X}[X]/\mathbb{E}_{Q_Y}[Y]$ en fonction de la vitesse à laquelle $\mathbb{E}_{Q_Y}[Y]$, $\mathbb{V}_{Q_Y}[Y]$, $\mathbb{E}_{Q_X}[X]$ et $\mathbb{V}_{Q_X}[X]$

sont autorisées à tendre vers zéro. Nous montrons ensuite que lorsque $\mathbb{E}_{Q_Y}[Y]$ (*resp.* $\mathbb{V}_{Q_Y}[Y]$) tend vers zéro suffisamment lentement (*resp.* rapidement), les intervalles de confiance basés sur le bootstrap nonparamétrique d'Efron ([67]) sont valides asymptotiquement au sens de (1.1). Ces résultats sont de nature asymptotique et nous les complétons à l'aide d'une approche complètement nonasymptotique. Pour ce faire, nous nous appuyons sur des résultats provenant de la littérature statistique tels que les inégalités de concentration (voir [25] pour une introduction) et les théorèmes d'impossibilité de [34]. Nous imposons des bornes supérieures sur les moments d'ordre 2 de $Q_{X,Y}$ ainsi qu'une borne inférieure sur $|E_{Q_Y}[Y]|$ strictement positive si bien que $|\mathbb{E}_{Q_X}[X]/\mathbb{E}_{Q_Y}[Y]|$ est borné par au-dessus uniformément sur le modèle et nous ne tombons donc pas dans le cadre de [63]. Etant données ces bornes, nous montrons comment construire des intervalles de confiance non asymptotiques pour chaque niveau de confiance en-dessous d'un seuil $\underline{t}_n$ et qui ont les propriétés suivantes: ils sont presque sûrement de longueur bornée et ils atteignent le niveau de confiance requis uniformément sur le modèle pour tout n fini. Les intervalles de confiance et $\underline{t}_n$ dépendent de n et des bornes sur les moments de $Q_{X,Y}$. Nous exhibons par ailleurs un niveau de confiance $\bar{t}_n$ au-dessus duquel il est impossible de construire un intervalle de confiance qui contienne $\overline{X}_n/\overline{Y}_n$ presque sûrement et qui est à la fois du niveau requis uniformément sur le modèle et presque sûrement de longueur bornée. Par conséquent, même en dehors du cadre de [63], une large classe d'intervalles de confiance incluant ceux basés sur la méthode delta ne peuvent pas être presque sûrement bornés et avoir un niveau de confiance garanti lorsque le niveau de confiance est trop proche de 1 pour un nombre d'observations n fini. Nous proposons un critère pour évaluer la fiabilité de la méthode delta en échantillon fini: quand il existe des bornes naturelles sur les moments de $Q_{X,Y}$, elles peuvent être utilisées pour calculer $\underline{t}_n$ pour avoir une idée du niveau de confiance maximum (qui dépend de n) auquel la méthode delta peut être employée de manière crédible pour construire des intervalles de confiance. Quand il n'existe pas de bornes naturelles, nous suggérons une règle du pouce: nous remplaçons les bornes par les moments empiriques correspondants. A l'aide de plusieurs exercices de simulation, nous recommandons de se baser sur $\underline{t}_n$ plutôt que sur $\bar{t}_n$ qui est trop conservateur. Nous présentons un autre résultat d'impossibilité quant à la longueur minimale qu'un intervalle de confiance uniformément valide peut avoir. Nous illustrons nos résultats asymptotiques et non asymptotiques à l'aide d'une application sur les disparités salariales liées au genre en France.

Résumé du chapitre 6

Ce chapitre propose un programme Stata qui implémente les différents outils statistiques introduits dans [53]. [53] part du cadre causal de Rubin et fait l'hypothèse que Z_o peut s'écrire $Z_o = DZ_o(1)+(1-D)Z_o(0)$. Une paire aléatoire (G,T) est également attribuée à chaque individu: T est la période ou la cohorte aléatoire à laquelle un individu appartient et G indique si un individu appartient à un groupe avec une intensité de traitement stable ou croissante entre périodes/cohortes. G identifie donc les groupes de traitement ($G=1$) et de contrôle ($G=0$) dans [53]. En notant S l'ensemble des individus du groupe de traitement qui passeraient de non-traités à traités s'ils étaient observés à plusieurs périodes, [53] donne plusieurs jeux d'hypothèses qui permettent d'identifier la quantité $\Delta = \mathbb{E}_{Q_{(Z_o(1),Z_o(0))|S,T=1}}[Z_o(1) - Z_o(0) \mid S, T=1]$ avec trois estimands différents. Le paramètre Δ est appelé un effet local de traitement moyen (LATE en anglais) et a été introduit dans [92]. Un des estimands appelé W_{DID} n'est pas nouveau et est très répandu en pratique tandis que les deux autres dénommés W_{TC} and W_{CIC} sont nouveaux. Sous les hypothèses d'identification qui sous-tendent le W_{CIC}, les auteurs montrent un résultat plus fort, à savoir que $Q_{Z_o(1)|S,T=1}$ et $Q_{Z_o(0)|S,T=1}$ sont identifiées de même que les effets locaux de traitement quantile (LQTEs en anglais) $\tau_\delta = q_{Q_{Z_o(1)|S,T=1}}(\delta) - q_{Q_{Z_o(0)|S,T=1}}(\delta)$. En sus de ces résultats d'identification, [53] propose des estimateurs pour les 4 estimands et prouvent leur normalité asymptotique. Un des

principaux enseignements de [53] est de montrer que les conditions requises pour identifier l'estimand très populaire qu'est le W_{DID} peuvent être assez peu vraisemblables dans certains cas. Le W_{TC} ainsi que le W_{CIC} peuvent alors être des alternatives utiles. Notre contribution revient à rendre les procédures d'estimation proposées par [53] disponibles sur le logiciel Stata qui est très utilisé en économétrie appliquée. En plus du calcul des estimateurs, nous construisons dans notre programme des intervalles de confiance à 95% sur Δ et τ_δ basés sur le bootstrap de même que des tests statistiques pour voir si les estimands de Δ sont significativement différents. L'inférence peut être rendue robuste à la dépendance en grappe unidimensionnelle. Dans [53] et [54], plusieurs extensions sont considérées: des versions modifiées de Δ sont définies pour traiter les cas où il y a plus de deux groupes, deux périodes et deux niveaux de traitement et des conditions suffisantes d'identification sont données; les résultats sont également étendus aux cas où les hypothèses d'identification ne sont valides que conditionnellement à un ensemble de covariables Z_e; des estimateurs adaptés sont proposés. Quand des covariables sont introduites dans le modèle, les estimateurs de W_{DID}, W_{TC} et W_{CIC} requièrent l'estimation de quantités du type $\mathbb{E}_{Q_{Z_o|G,T,X}}[Z_o \mid G,T,X]$ et $\mathbb{E}_{Q_{D|G,T,X}}[D \mid G,T,X]$. [54] prouve la normalité asymptotique des estimateurs de W_{DID}, W_{TC} et W_{CIC} quand les espérances conditionnelles précédentes sont estimées nonparamétriquement à l'aide de régression polynomiales. Notre commande Stata propose aussi ces estimateurs. Les espérances conditionnelles $\mathbb{E}_{Q_{Z_o|G,T,X}}[Z_o \mid G,T,X]$ et $\mathbb{E}_{Q_{D|G,T,X}}[D \mid G,T,X]$ peuvent être estimées par moindres carrés ordinaires, Probit ou Logit (lorsque Z_o ou D est binaire) ou régression polynomiale nonparamétrique. L'ordre de la régression polynomiale peut être spécifié par l'utilisateur ou choisi automatiquement par validation croisée basée sur le critère de l'erreur quadratique moyenne (voir [135] pour des définitions). De la même manière que [54], nous revisitons l'article empirique de [76] pour montrer comment utiliser notre commande Stata et pour mettre en exergue les différences qui apparaissent lorsque l'on estime le W_{DID} plutôt que le W_{TC} par exemple. Nous concluons avec un exercice de simulation substantiel pour vérifier la performance de nos estimateurs dans des échantillons de taille modérée. Pour chaque modèle choisi pour simuler les données, nous lançons 1000 réplications de ce modèle, à chaque fois pour 3 tailles d'échantillon différentes, à savoir 400, 800 et 1600. Nous évaluons la qualité de nos estimateurs à l'aide du biais moyen, de la moyenne de l'erreur quadratique moyenne et du taux de couverture estimé. Les moyennes et le taux de couverture sont calculés sur les 1000 réplications.

Chapter 2

Introduction in English

What is the connection between this PhD dissertation and real-life economic problems? This concern seems supported by the somewhat technical title of the here-presented work. These introductory words are mainly devoted to addressing this topic.

How do low-income people allocate their budget between luxury goods such as branded clothing and food? Do unemployed workers have higher chances of returning to work when they get trained by job centers? These are two examples of questions economists are interested in ([46, 47]). In the first example, the goal is to understand the mechanisms behind individual economic decisions. In the second case, interest lies in measuring the impact of a public policy. To answer these questions, economic theory provides predictions that have to be tested against observational data. To test economic predictions, some restrictions have to be imposed on how observations are generated. These constraints form a model of observed behaviours. It is implausible to assert we can explain perfectly consumption (*resp.* return to work) in terms of budget allocation (*resp.* training expenditures). It is more sensible to assume that consumption or return to work depend also on unobserved factors that capture complex and indescribable phenomena. When the observed and unobserved components of the problem are treated as random, we are left with a statistical model. As discussed in the introductory chapter of [60], statistics is the generic interface between theories we want to test and data. Our work lies in a discipline called econometrics. The latter is a subfield of economics that uses statistical tools to adress socio-economic questions. In this introduction, we aim at comparing econometrics with different subfields of statistics and in particular statistical learning. Statistical learning is a discipline that studies theoretical properties of machine learning algorithms when the data is supposed to be generated according to a statistical model. As will be emphasized, econometrics and statistical learning somehow differ in their definition of a statistical model. Note further that depending on how general a statistical model is, it will be called parametric, semiparametric or nonparametric. We give precise definitions of the latter notions in the rest of this introduction.

Key notions in semi- and nonparametric statistics

This dissertation is concerned with semi- and nonparametric statistical models. To explain accurately those notions and understand them better, a few definitions are in order. We consider a random vector W mapping an underlying probability space $(\Omega, \mathcal{A}, P)$ to a measurable space $(E, \mathcal{E})$. W refer to all the random components of the model, be they observed or not. We assume that E can be given a metric space structure thanks to the norm $||\cdot||_E$. $\mathcal{Q}$ corresponds to the set of all probability distributions defined on $(E, \mathcal{E})$. In this work, we always consider that the distribution of W denoted by Q_W belongs to a strict subset of $\mathcal{Q}$ that we call $\mathcal{Q}^*$. A prominent example is $\mathcal{Q}^* := \left\{Q \in \mathcal{Q} : \mathbb{E}_Q\left[||W||_E^2\right] < +\infty\right\}$, where $\mathbb{E}_Q$ is

the expectation operator under Q. This example is a nonparametric subset of $\mathcal{Q}$ since elements in $\mathcal{Q}^*$ are not fully characterized by a finite-dimensional parameter. We actually only consider nonparametric subsets of $\mathcal{Q}$ here. A statistical model is constructed by: i) choosing a set Θ called the parameter set; ii) associating to each $\theta \in \Theta$ a distribution $Q_\theta \in \mathcal{Q}^*$. To fix ideas we give the example of the canonical linear regression model inspired by the introductory chapter of [60]: $Z_o = Z_e'\beta + \epsilon$, where $Z_e \in \mathbb{R}^p$. The unusual notations Z_o for the outcome variable and Z_e for the explanatory vector are introduced to be consistent with latter chapters. We let $Z = (Z_o, Z_e')'$. In this example, the parameter is $\theta = (\beta, Q_{Z_e,\epsilon})$. The parameter set is $\Theta = \mathbb{R}^p \times \mathcal{D}$ with $\mathcal{D} := \left\{Q : \mathbb{E}_Q[Z_e\epsilon] = 0, \mathbb{E}_Q[Z_eZ_e']^{-1} < +\infty\right\}$ and $\mathcal{Q}^* = \left\{Q : \mathbb{E}_Q[Z_eZ_e']^{-1} < +\infty\right\}$. In the linear regression model, we are only interested in β which can be formally written $\beta = T(\theta)$ for T a projection map. It is often the case that the parameter of interest is not θ itself but some transformation of it. When $T(\theta)$ is a finite-dimensional quantity we call the model semiparametric, otherwise we call it nonparametric.

A fundamental question amounts to asking whether a statistical model is identified: a model is called identified if every $Q \in \mathcal{Q}^*$ can be generated by at most one $\theta \in \Theta$. In what follows, we assume that model identification holds. Exhibiting primitive identification conditions is in general far from trivial but lies outside the scope of this manuscript. Readers should keep in mind that we focus mostly on models for which identification is (relatively) well-understood. In the linear regression case, the restrictions $\mathbb{E}_Q[Z_eZ_e']^{-1} < +\infty$ and $\mathbb{E}_Q[Z_e\epsilon] = 0$ are necessary and sufficient for identification for instance.

When a model is identified, its parameters can be expressed in terms of the distribution of observable random variables ([60]). Given identification, the main task of an econometrician is to use observations to estimate $T(\theta)$ and conduct inference on this quantity. From now on, we assume we have at our disposal n observations $(Z_i)_{i=1}^n$ with joint distribution Q_n. We impose that all observations have the same marginal distribution Q_Z, *i.e* be identically distributed. We further restrict $T(\theta)$ to live in a metric space $(\mathcal{T}, ||\cdot||_T)$. An estimator $\widehat{T}(\theta)$ is a measurable function of $(Z_i)_{i=1}^n$ that takes its values in $\mathcal{T}$. The quality of an estimator is measured by $\left|\left|\widehat{T}(\theta) - T(\theta)\right|\right|_T$. An estimator is consistent when $\left|\left|\widehat{T}(\theta) - T(\theta)\right|\right|_T$ goes to 0 in probability as n increases. In general, the choice of the norm $||\cdot||_T$ is not unique. When $\mathcal{T}$ is finite-dimensional, this choice is not crucial as all norms are equivalent. On the other hand, when $\mathcal{T}$ is infinite-dimensional, norms are not all equivalent anymore. It can then be the case that an estimator is consistent for one norm and not for another one. In infinite-dimensional problems, the discrepancy between different norms can in fact be very useful: it is sometimes possible to use one norm as a regularization tool to help obtain consistency of an estimator with respect to the other one. The notion of statistical regularization is explained in more details below. Inference covers two closely connected topics: confidence intervals and hypothesis tests. Since we only focus on confidence sets in the next chapters, we omit the definition of hypothesis tests here. What follows is largely based on Chapter 6 in [79]. A confidence set (CS) is loosely speaking a random subset C_n of $\mathcal{T}$ that depends on $(Z_i)_{i=1}^n$ but not on $T(\theta)$. We present the asymptotic criteria that are used to assess the quality of a CS. Given $\delta \in (0,1)$, a confidence set has asymptotic level $1-\delta$ pointwise over Θ if

$$\inf_{\theta\in\Theta} \liminf_{n\to+\infty} \mathbb{P}_{Q_{\theta,n}}(C_n \ni T(\theta)) \geq 1-\delta, \tag{2.1}$$

and it has asymptotic level $1-\delta$ uniformly over Θ if

$$\liminf_{n\to+\infty} \inf_{\theta\in\Theta} \mathbb{P}_{Q_{\theta,n}}(C_n \ni T(\theta)) \geq 1-\delta. \tag{2.2}$$

The second criterion ([104, 29]) that is sometimes called the honesty criterion is obviously more demanding than the first one and has raised a lot of attention especially in the nonparametric statistics community. Those first two criteria ensure that the CS is in some sense asymptotically “reliable". They

are not sufficient though as they do not throw away noninformative CS such as picking $C_n = \mathcal{T}$ for every n. A second requirement is thus that the CS satisfies some optimality rule. Optimality can de defined in several ways. A CS can be said pointwise/uniformly optimal if the inequality in (2.1)/(2.2) becomes an equality. Another popular rule asks for the diameter of the CS to shrink to zero in probability sufficiently fast as n goes to infinity. To discriminate between confidence sets that verify the previous optimality criteria, one can for instance study the limit of the ratio of their diameters.

The statistical learning paradigm of Regularized Empirical Risk Minimization (RERM)

To see the connections and differences between econometrics and statistical learning, we first need to understand the general goal of statistical learning and the theoretical framework that results from this goal. We stick to the case where the observed data can be divided into two parts: an outcome $Z_o \in \mathcal{Z}_o$ and a set of potential predictors $Z_e \in \mathcal{Z}_e$, with joint distribution Q_{Z_o,Z_e}. We still use the notation $Z = (Z_o, Z_e')'$. The aim is to predict Z_o as accurately as possible using a function of Z_e according to a rule that captures the quality of the prediction performance. The prediction rule (also called loss) and class of functions are chosen by the researcher and those choices are largely driven by computational considerations. Put formally, the theoretical problem is: given a class of functions $\mathcal{H}$ mapping $\mathcal{Z}_e$ to $\mathcal{V}$, and a loss $\ell : \mathcal{Z}_o \times \mathcal{Z}_e \times \mathcal{H} \mapsto \mathbb{R}^+$, it is assumed that there exists $h^* \in \mathcal{H}$ not necessarily unique such that

$$h^* \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} \, \mathbb{E}_{Q_Z} \left[\ell(Z_o, Z_e, h) \right]. \tag{2.3}$$

For instance, we could take $\ell(z_o, z_e, h) = (z_o - h(z_e))^2$, *i.e* the least-squares loss, and $\mathcal{H} := \left\{ h : \mathcal{Z}_e \to \mathbb{R} \text{ s.t } \sup_{z_e \in \mathcal{Z}_e} |h(z_e)| \leq M \right\}$. This amounts to solving a nonparametric least-squares problem. Note that in econometrics, the nonparametric least-squares problem is only of interest if we assume that the data are generated according to the model $Z_o = h^{**}(Z_e) + \epsilon$, subject to $\mathbb{E}_{Q_{\epsilon | Z_e}}[\epsilon \mid Z_e] = 0$. As a matter of fact, if $h^{**} \in \mathcal{H}$, it is a standard fact that h^{**} satisfies (2.3). What happens when $h^{**} \notin \mathcal{H}$? The problem in (2.3) is still well-defined and admits a solution but this solution is not h^{**} and is suboptimal: we have $\mathbb{E}_{Q_Z}[\ell(Z_o, Z_e, h^{**})] < \mathbb{E}_{Q_Z}[\ell(Z_o, Z_e, h^*)]$. In econometrics terms, solving (2.3) when $h^{**} \notin \mathcal{H}$ is equivalent to focusing on a misspecified model. In the statistical learning framework, this misspecification is in general allowed. The statistical learning paradigm has other particularities: interest lies mostly in high-dimensional models in which the class of functions $\mathcal{H}$ is allowed to grow larger with n. In a typical high-dimensional model, $h(Z_e)$ takes the form $Z_e'\beta$ where $\beta \in \mathbb{R}^p$ and p is potentially much larger than n. To make the problem solvable, a common assumption is that of sparsity, *i.e* only s entries (with s small relative to n) are nonzero in the vector β. A generalization of sparsity called approximate sparsity is also popular: it imposes that β can be well approximated (but not necessarily fully recovered) by a small number of its entries. Approximate sparsity bears strong ties with classical nonparametric models and is close to the notion of smoothness of a function. We refer to [15], [16] and [17] for enlightening discussions on this question. How can h^* be recovered from observations? The n observations in the sample $(Z_i)_{i=1}^n$ are assumed to be independent and identically distributed (*i.i.d*) and a naive approach would consider the direct empirical counterpart of (2.3): $h_n \in \operatorname{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(Z_{o,i}, Z_{e,i}, h)$. This is however not satisfactory in a high-dimensional setup. As a matter of fact, the previous empirical minimization problem does not take advantage of the sparsity assumption at all. To circumvent this, what matters is to endow $\mathcal{H}$ with a norm $||\cdot||_R$ which captures well sparsity and use this norm to regularize the empirical minimization procedure. The problem becomes

$$h_n \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(Z_{o,i}, Z_{e,i}, h) + \alpha \, ||h||_R^p \right\}, \tag{2.4}$$

and is called the Regularized Empirical Risk Minimization procedure (RERM). The quantity α is the weight put on the regularization term and its choice is key to get theoretical results on h_n. The exponent p is chosen equal to 1 or 2 most of the time. The most famous procedure that fits in this general framework is the Lasso ([133]) for linear regression: $\ell(Z_{o,i}, Z_{e,i}, h) = (Z_{o,i} - Z'_{e,i}\beta)^2$, $\mathcal{H} = \{\langle \cdot, \beta \rangle, \beta \in \mathbb{R}^p\}$, $p = 1$ and $||h||_R = ||\beta||_1$ where $||\cdot||_1$ is the ℓ_1 norm in $\mathbb{R}^p$. The quality of the estimated minimizer h_n is measured by the so-called excess risk criterion $\mathcal{R}(h_n) := \mathbb{E}_{Q_Z}\left[\ell(Z_o, Z_e, h_n) \mid (Z_i)_{i=1}^n\right] - \mathbb{E}_{Q_Z}\left[\ell(Z_o, Z_e, h^*)\right]$. In the previous definition, (Z_o, Z_e) is an independent copy of the sequence $(Z_{o,i}, Z_{e,i})_{i=1}^n$. In statistical learning, the goal is to control the probability that the excess risk is larger than an explicit threshold for a fixed number of observations. This is called an oracle prediction inequality and its general form is: for every $\delta \in (0,1)$ and every $n \geq 1$, $\mathbb{P}_{Q_{(Z_i)_{i=1}^n}}(\mathcal{R}(h_n) > \gamma(n,\delta)) < \delta$. The function γ may depend on ℓ, $\mathcal{H}$, Q_Z and some universal constants and for a fixed δ $\gamma(n,\delta)$ is decreasing in n. Sometimes, the results are weaker in that they may not hold for every $\delta \in (0,1)$ and may require n to be larger than some threshold. Even though prediction is of particular importance, a strand of the statistical learning community is also interested in the estimation properties of h_n. The chosen criterion is $||h_n - h^*||_H$ for a norm that usually differs from the regularization norm (many examples of $||\cdot||_H$ and $||\cdot||_R$ are given in [5]). To derive an estimation result, we note that h^* should be unique or at least it should be possible to uniquely select one of the minimizers of the problem (2.3). Over the past twenty years, several conditions have been proposed to relate $||h_n - h^*||_H$ and $\mathcal{R}(h_n)$ and thus directly obtain an estimation oracle inequality from the prediction oracle ([106, 134, 5, 45])

$$\mathbb{P}_{Q_{(Z_i)_{i=1}^n}}(||h_n - h^*||_H > \gamma(n,\delta)) < \delta, \quad \forall(\delta, n) \in (0,1) \times \mathbb{N}^*. \tag{2.5}$$

The econometric interpretation of RERM and the need for other tools to handle endogeneity

Many econometric models can actually be written using the RERM framework discussed in the previous paragraph but the motivation for using RERM differs from that in statistical learning. We illustrate this point focusing on mean and median regression models, *i.e* we assume that the data are generated according to $Z_o = h^{**}(Z_e) + \epsilon$ with either the restriction $\mathbb{E}_{Q_{\epsilon|Z_e}}[\epsilon \mid Z_e] = 0$ or $med(Q_{\epsilon|Z_e}) = 0$. Under the first restriction, h^{**} is the true mean regression function of Z_o given Z_e and satisfies $h^{**} = \operatorname{argmin}_{h:\mathbb{E}_{Q_{Z_e}}[h(Z_e)^2]<+\infty} \mathbb{E}_{Q_Z}[(Z_o - h(Z_e))^2]$. Under the second restriction, h^{**} is the true median regression function of Z_o given Z_e and satisfies $h^{**} = \operatorname{argmin}_{h:\mathbb{E}_{Q_{Z_e}}[|h(Z_e)|]<+\infty} \mathbb{E}_{Q_Z}[|Z_o - h(Z_e)|]$. Economic theory often provides natural constraints on h^{**} such as monotonicity, convexity/concavity or smoothness. Let $\mathcal{C}$ denote the set of all measurable functions from $\mathcal{Z}_e$ to $\mathcal{Z}_o$ that satisfy the economic-related constraints. The class of functions can then be chosen as $\mathcal{H} = \left\{h : \mathbb{E}_{Q_{Z_e}}[h(Z_e)^2] < +\infty\right\} \cap \mathcal{C}$ in the mean regression case or $\mathcal{H} = \left\{h : \mathbb{E}_{Q_{Z_e}}[|h(Z_e)|] < +\infty\right\} \cap \mathcal{C}$ in the median regression case. The mean regression case fits the RERM framework with $\ell(Z_o, Z_e, h) = (Z_o - h(Z_e))^2$ and the median regression boils down to choosing $\ell(Z_o, Z_e, h) = |Z_o - h(Z_e)|$. What differs between econometrics and statistical learning is the fact that ℓ is dictated by the parameter of interest in econometrics and is not chosen: if one wants to recover the true mean regression function, $\ell(\cdot,\cdot,\cdot)$ is necessarily the least-squares loss. What is more, $\mathcal{H}$ is chosen to capture some economically justified restrictions, not only on computational grounds. To make the connection even clearer, it is useful to remark that (2.3) and (2.4) are equivalent to the class of regularized M-estimators, a name that is perhaps more familiar in econometrics.

The RERM is a very general setting which is however not very well-suited to deal with one notion that is central in econometrics: endogeneity. This concept captures the idea that certain variables that influence both the outcome Z_o and the observed explanatory variables Z_e may not be observable by the econometrician. In that case, using solely Z_e to explain Z_o does not allow to recover the parameters of

interest in general. To overcome this issue, a standard approach is to find additional variables called instruments that have an impact on Z_o only through the explanatory vector Z_e. We do not explain here the formal reasons behind the lack of compatibility between endogeneity and M-estimation. Intuitively, the M-estimator formulation is based on a projection argument that does not combine well with techniques to remove endogeneity (outside of linear models at least, see [62]). Under endogeneity, it is in fact more natural to obtain the parameters of interest by finding the zero of an appropriate set of moment conditions ([37]). To see this, we focus on the median regression case. We now assume that the model is $Z_o = h^{**}(Z_e) + \epsilon$ with $med(Q_{\epsilon|Z_e}) \neq 0$ but $med(Q_{\epsilon|X}) = 0$. In this model, at least one of the components of Z_e is linked to ϵ which is why the quantile restriction $med(Q_{\epsilon|Z_e}) = 0$ breaks down. The vector X consists of all the extra instruments plus the variables in Z_e that do not violate the quantile restriction. It is possible to show that the median regression model can be expressed as

$$\mathbb{E}_{Q_{Z|X}}\left[\mathbb{1}\left\{Z_o \leq h(Z_e)\right\} \mid X\right] = 0 \;\; Q_X - a.s \iff h = h^{**}.$$

This example motivates the following general class of problems as an alternative to the M-estimator framework:

$$\mathbb{E}_{Q_{Z|X}}\left[\rho(Z,h) \mid X\right] = 0 \;\; Q_X - a.s \iff h = h^{**}, \tag{2.6}$$

where ρ is a known finite-dimensional vector of functions. The relation in (2.6) can be equivalently written as $h^{**} = \operatorname{argmin}_{h\in\mathcal{H}} \mathbb{E}_{Q_X}\left[\|\mathbb{E}_{Q_{Z|X}}\left[\rho(Z,h) \mid X\right]\|^2\right]$, with $\|\cdot\|$ the Euclidean norm. If some natural constraints can be imposed on h^{**}, it is useful to endow $\mathcal{H}$ with a regularization norm $\|\cdot\|_R$ that magnifies these constraints. The empirical analogue of the problem becomes

$$h_n \in \underset{h\in\mathcal{H}}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \|\mathbb{E}_n\left[\rho(Z,h) \mid X = X_i\right]\|^2 + \alpha\|h\|_R^2 \right\}. \tag{2.7}$$

The quantity $\mathbb{E}_n\left[\rho(Z,h) \mid X = \cdot\right]$ stands for any estimator of the function $\mathbb{E}_{Q_{Z|X}}\left[\rho(Z,h) \mid X = \cdot\right]$. The estimation procedure in (2.7) is called the Regularized Generalized Method of Moments (RGMM) approach. When the class of functions $\mathcal{H}$ is parametric, it is in general useless to add a regularization term. When $\mathcal{H}$ is parametric, h^{**} can even be identified using a finite number of unconditional moment conditions in some cases, *i.e* $\mathbb{E}_{Q_Z}\left[\rho(Z,h)\right] = 0 \iff h = h^{**}$. In this simplified setup, there is a natural connection between M-estimators and GMMs. To explain this connection, we assume for simplicity that $h(Z_e) = Z_e'\beta$. Under some conditions on $\mathbb{E}_{Q_Z}\left[\ell(Z_o, Z_e, \langle\cdot,\beta\rangle)\right]$ including differentiability in β, (2.3) is equivalent to

$$\frac{\partial}{\partial\beta}\mathbb{E}_{Q_Z}\left[\ell(Z_o, Z_e, \langle\cdot,\beta\rangle)\right] = 0 \iff \beta = \beta^{**}.$$

If there exists a function $\rho : (z_o, z_e, \beta) \mapsto \rho(z, \langle\cdot,\beta\rangle)$ such that for every β $\frac{\partial}{\partial\beta}\mathbb{E}_{Q_Z}\left[\ell(Z_o, Z_e, \langle\cdot,\beta\rangle)\right] = \mathbb{E}_{Q_Z}\left[\rho(Z, \langle\cdot,\beta\rangle)\right]$, then the M-estimation problem has been turned into a GMM one. GMMs that are obtained from the first-order condition of a M-estimator are called Z-estimators (this is explained in Chapter 5 of [136]).

The general framework of (2.7) has been investigated in many contributions in the econometric literature, a landmark being [37]. Unlike the statistical learning approach, the prediction properties of h_n are not of central importance. Most results consist in proving that $\|h_n - h^{**}\|_H$ converges to 0 at a fast enough rate, for a norm $\|\cdot\|_H$ possibly different from $\|\cdot\|_R$. Results are asymptotic most of the time, in the sense that estimation oracles valid for every n are usually not exhibited.

The construction of confidence sets in econometrics and nonparametric statistics

In principle, confidence sets could be computed without resorting to an estimator of the parameter of interest. In practice however, CSs are almost systematically built based on an estimator and there is de facto a strong connection between estimation and the construction of CSs. To understand this connection, we discuss two cases: the construction of a CS for h^{**} in the model (2.3) (with $h^* = h^{**}$) and the construction of a CS for a functional of h^{**} denoted $\varphi(h^{**})$.

The first example is typically what would be of interest in nonparametric statistics ([79]). If an oracle inequality similar to (2.5) exists and $\gamma(\cdot,\cdot)$ does not depend on h^{**}, a CS of level $1-\delta$ valid for every n uniformly over $\mathcal{H}$ can be constructed by collecting every $h \in \mathcal{H}$ such that $\|h_n - h\|_H \leq \gamma(n,\delta)$. This is an appealing theoretical finding since it applies to a wide range of statistical problems but its direct implementation is often difficult: the norm $\|\cdot\|_H$ can be cumbersome to compute and the search for all the hs that fall in the CS may be computationally demanding; the function $\gamma(\cdot,\cdot)$ depends in general on constants that are either unkown, very large or must be estimated; even when the two previous difficulties do not arise, CSs based on an oracle inequality may have a diameter that is asymptotically too large in a sense made precise below. To underline those challenges, we discuss a very simple problem. We want to construct a confidence interval for $\mathbb{E}_{Q_{Z_o}}[Z_o]$ based on n *i.i.d* draws $(Z_{o,i})_{i=1}^n \sim Q_{Z_o}$. We assume that the variance of Q_{Z_o} is finite and known equal to V. An application of the Bienaymé-Chebyshev inequality (see Chapter 2 in [136]) yields the following oracle inequality: for every $n \geq 1$ and $\delta \in (0,1)$, $\mathbb{P}_{Q_{Z_o}^{\otimes n}}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_{o,i} - \mathbb{E}_{Q_{Z_o}}[Z_o]\right| > \sqrt{\frac{V}{n\delta}}\right) < \delta$. The interval $I_n^{\delta,1} := \left[\frac{1}{n}\sum_{i=1}^n Z_{o,i} - \sqrt{V/n\delta}, \frac{1}{n}\sum_{i=1}^n Z_{o,i} + \sqrt{V/n\delta}\right]$ verifies for every $n \geq 1$

$$\inf_{\mathbb{E}_{Q_{Z_o}} \in \mathbb{R}} \mathbb{P}_{Q_{Z_o}^{\otimes n}}\left(I_n^{\delta,1} \ni \mathbb{E}_{Q_{Z_o}}[Z_o]\right) \geq 1-\delta.$$

Assuming that the variance is known equal to V is unfortunately simplistic and applied statisticians would not be ready to impose this. A solution would be to: i) assume that the true variance is unknown but upper bounded by V which would yield the same result as before, ii) replace V by an estimator of the variance but the nonasymptotic guarantees associated with $I_n^{\delta,1}$ would collapse. Note that i) is often not appealing in practice as it is hard to come up with a sensible value for the upper bound V. This restriction (or bounds on higher-order moments) is however unavoidable to conduct nonasymptotic inference.

The second example is central to econometrics where the parameter of interest is often not h^{**} as defined in (2.6) but a scalar transformation thereof (see the introduction of Chapter 3 and references therein for several concrete illustrations). In econometrics, the approach to constructing CSs is mainly asymptotic: the standard approach relies on exhibiting a suitable (random) sequence r_n such that the distribution of $r_n(\varphi(h_n) - \varphi(h^{**}))$ converges to a $\mathcal{N}(0,1)$ distribution. Let $q_{\mathcal{N}(0,1)}(1-\delta/2)$ stand for the $1-\delta/2$ quantile of the $\mathcal{N}(0,1)$ distribution. The interval $I_n^{\delta,2} := \left[\varphi(h_n) - q_{\mathcal{N}(0,1)}(1-\delta/2)/r_n, \varphi(h_n) + q_{\mathcal{N}(0,1)}(1-\delta/2)/r_n\right]$ can be shown to be pointwise of level $1-\delta$ asymptotically. It satisfies the optimality criteria introduced earlier, in particular the probability that $\varphi(h^{**})$ belongs to $I_n^{\delta,2}$ tends to $1-\delta$ for every $\delta \in (0,1)$ and $h^{**} \in \mathcal{H}$. The main drawback of $I_n^{\delta,2}$ is that its behaviour is uncontrolled for every finite n and it is not honest as defined previously without further restrictions (see [97]).

Coming back to the construction of a confidence interval for $\mathbb{E}_{Q_{Z_o}}$ with a known variance V, we can build a pointwise asymptotically valid confidence interval as $I_n^{\delta,3} := \left[\frac{1}{n}\sum_{i=1}^n Z_{o,i} - q_{\mathcal{N}(0,1)}(1-\delta/2)\sqrt{V/n}, \frac{1}{n}\sum_{i=1}^n Z_{o,i} + q_{\mathcal{N}(0,1)}(1-\delta/2)\sqrt{V/n}\right]$. If we compute the ratio of the lengths of $I_n^{\delta,1}$ and $I_n^{\delta,3}$ and study its limit in probability, we remark that $diam(I_n^{\delta,1})/diam(I_n^{\delta,3}) \to 1/(q_{\mathcal{N}(0,1)}(1-\delta/2)\delta)$ which can be shown to be larger than 1 for every $\delta \in (0,1/2)$. This implies that $I_n^{\delta,1}$ is asymptotically of level strictly larger than $1-\delta$ and is therefore conservative.

The previous discussion highlights the fact that it is difficult to combine asymptotic optimality and honesty. Optimality and honesty are not incompatible though and a vast literature tackling this question has blossomed ([87, 123, 122]). In recent years, several econometricians have been active in this field and proposed methods both theoretically appealing and practical ([11, 10]).

Relaxing the *i.i.d* assumption matters in econometrics

There are many natural reasons to go beyond the *i.i.d* assumption. Perhaps the most pervasive one is time: when time plays a role, which is the case in time series or panel data, data is allowed to be dependent over time and to have a time-varying distribution. As a result, observations are neither independent nor identically distributed anymore. We do not discuss the issue of time in statistical modelling any longer since time is never a core element of the models we study in this dissertation.

Even in the context of cross-sectional data (data that is not indexed by time), *i.i.d*-ness is often deemed implausible by applied econometricians. Let us take a simple example: we observe a sample of n workers and we have information on their commuter zone and industry. It is quite standard to allow for unobserved aggregate economic shocks at the geographical area and industry levels ([1, 27, 110]). The goal is to build CSs that are *robust* to the presence of such shocks. CSs are called robust if they have (asymptotic) coverage at the desired level should the data be *i.i.d* or not. The *i.i.d* assumption is also not very credible with interaction data, that is data that stems from the interactions of the individuals of one population among themselves. In this setting, datasets have the form $(W_{i,j})_{1\leq i\neq j\leq n}$ where $W_{i,j}$ is an observation relative to the pair formed by individuals i and j. Those notions of cross-sectional dependence exist in other statistical fields such as spatial statistics or network analysis. In those fields however, dependence tends to be the main topic of interest, *i.e* a model on the dependence structure is formed and the goal is to recover the parameters of the former. In econometrics (or part of it at least), the aim is quite different: dependence is mainly seen as a nuisance term that has to be accounted for to conduct valid inference on some other quantity. Cross-sectional dependence is at the heart of Chapter 4.

In the preceding paragraph we do not relax the assumption that observations are identically distributed. We never give up on that assumption in that dissertation and we view it as quite fundamental (except in the case of data that exhibit a time dimension): as a matter of fact, it seems farily natural to assume that two individuals from the same sample - no matter how different they may be in terms of education and wage for instance - are simply two distinct draws from the same distribution. Some researchers have a different view on the matter: they take the observed explanatory variables $(Z_{e,i})_{i=1}^n$ as fixed and nonrandom which leads to a non identically distributed sample (see Chapter 2.8 in [136]).

Causality and machine learning

Causality is one of the pillars of the econometric discipline. This notion became popular in econometrics following an article by Donal Rubin ([124]). It relies on a thought experiment: there exist two states of the nature (labelled 0 and 1) and each individual is placed in one of the two. Individuals are given an outcome variable $Z_o(0)$ or $Z_o(1)$ depending on which state they are in. At the individual level, the causal impact of changing states simply is the difference $Z_o(1) - Z_o(0)$. Why is causality interesting in econometrics? It is a convenient framework to model the impact of a public policy at the aggregate level. If the government could observe $Z_o(1) - Z_o(0)$ for everybody, this government could measure the consequence of making people switch states according to some predefined criterion. In this context, enforcing a public policy is equivalent to making individuals switch states.

In reality, the government observes either $Z_o(1)$ or $Z_o(0)$ but never both: the causal framework is an example of a missing data statistical problem ([121]). Denoting D the state individuals are in, the government only observes $Z_o = DZ_o(1) + (1 - D)Z_o(0)$. Without further restrictions, it is only possible

to recover $Q_{Z_o(1)|D=1}$ and $Q_{Z_o(0)|D=0}$. Imposing further $(Z_o(1), Z_o(0)) \perp\!\!\!\perp D$ ensures that $Q_{Z_o(1)|D=1} = Q_{Z_o(1)}$ and $Q_{Z_o(0)|D=0} = Q_{Z_o(0)}$. We refer to [81] for a thorough presentation of the identification question in Rubin's causal framework. Identifying $Q_{Z_o(0)}$ and $Q_{Z_o(1)}$ allows to compute the average change associated with the treatment D: $\mathbb{E}_{Q_{Z_o(0),Z_o(1)}}[Z_o(1)-Z_o(0)]$, or the change in the δ-th quantile: $q_{Q_{Z_o(1)}}(\delta) - q_{Q_{Z_o(0)}}(\delta)$. On the other hand, it does not allow to recover the δ-th quantile of the treatment effect $q_{Q_{Z_o(1)-Z_o(0)}}(\delta)$. To get $q_{Q_{Z_o(1)}}(\delta) - q_{Q_{Z_o(0)}}(\delta) = q_{Q_{Z_o(1)-Z_o(0)}}(\delta)$, one has to assume that the rank of an individual under $Q_{Z_o(0)}$ is the same under $Q_{Z_o(1)}$ (rank invariance property, *cf* [65]).

In the remaining of this paragraph, we focus on the parameter $\mathbb{E}_{Q_{Z_o(0),Z_o(1)}}[Z_o(1) - Z_o(0)]$ which we denote by τ. One drawback of the assumption $(Z_o(1), Z_o(0)) \perp\!\!\!\perp D$ is its non-testability. It is often replaced with $(Z_o(1), Z_o(0)) \perp\!\!\!\perp D \mid Z_e$ which is not testable either but strictly weaker. Under this last assumption, one can show ([81]) $\tau = \mathbb{E}_{Q_{Z_e}}\left[\mathbb{E}[Z_o \mid D = 1, Z_e] - \mathbb{E}[Z_o \mid D = 0, Z_e]\right]$. The right-hand side depends only on observable variables. The two tasks researchers are mainly interested in are i) estimation of and inference on τ, ii) testing for heterogeneity of the treatment effect for different individual profiles z_e. This second goal consists in testing whether $\mathbb{E}_{Q_{(Z_o(0),Z_o(1))|Z_e}}[Z_o(1) - Z_o(0) \mid Z_e = z_1] = \mathbb{E}_{Q_{(Z_o(0),Z_o(1))|Z_e}}[Z_o(1) - Z_o(0) \mid Z_e = z_2]$ when $z_1 \neq z_2$. In both cases, a first step consists in estimating the functions $\mathbb{E}[Z_o \mid D = 1, Z_e = \cdot]$ and $\mathbb{E}[Z_o \mid D = 0, Z_e = \cdot]$ (only evaluated at points z_1 and z_2 in the second case). How to estimate those functions in a flexible fashion? One possibility is to use classical nonparametric tools such as Nadaraya-Watson or local linear regressions ([135]). The theoretical guarantees of these methods have been long established ([59, 70]). Their main limitation is their poor performance in practice when the dimension of Z_e is large. On the other hand, machine learning techniques such as random forests or deep neural networks perform well on simulations even when the dimension of Z_e is large but their theoretical properties are much less known. Recent efforts both from the econometrics and statistical learning communities have led to theoretical advances on machine learning algorithms: Theorem 3 in [71] shows the asymptotic normality of an estimator of τ based on a deep neural network architecture, [138] prove the asymptotic normality of a random forest method to estimate $\mathbb{E}_{Q_{(Z_o(0),Z_o(1))|Z_e}}[Z_o(1) - Z_o(0) \mid Z_e = z_e]$ for a fixed z_e. Quite interestingly, the theoretical properties are not very different from those of more classical nonparametric tools: deep neural networks have been shown to work for exactly the same functions as more classical nonparametric tools and suffer from the same curse of dimensionality in the Z_e vector; random forests can approximate functions that are less smooth than standard methods but are still subject to the curse of dimensionality.

Summary of Chapter 3

In this chapter, we focus on the generic problem (2.6). As was explained before, many research articles (actually most) interested in this problem build an estimator based on (2.7) ([3], [112], [20], and [37] to name a few). There are actually other possibilities to construct an estimator for this class of problems and we look at the family of Generalized Empirical Likelihood (GEL) estimators ([113], [99]). To present GEL estimators, it is easier to start with a simplified version of (2.6): we assume that h is replaced with a finite-dimensional parameter $\beta \in \mathcal{B}$ and the true value β^{**} is such that $\mathbb{E}_{Q_Z}[\rho(Z, \beta)] = 0 \iff \beta = \beta^{**}$. As explained in [113, 99], β^{**} can equivalently be identified by

$$\beta^{**} = \underset{\beta \in \mathcal{B}}{\operatorname{argmin}} \sup_{\lambda \in \Lambda(\beta, Q_Z)} \mathbb{E}_{Q_Z}\left[\psi_\gamma(\lambda' \rho(Z, \beta))\right], \tag{2.8}$$

where $\Lambda(\beta, Q_Z) := \bigcap_{z \in supp(Q_Z)} \{\lambda : \psi_\gamma(\lambda' \rho(z, \beta)) \text{ exists}\}$ and $\psi_\gamma : u \mapsto \frac{2}{\gamma}\left[-(\gamma+1)\frac{u+1}{2}\right]^{\frac{\gamma}{\gamma+1}} - \frac{2}{\gamma(\gamma+1)}$. Taking the sample analogue of the previous saddle point problem yields one estimator for each function ψ_γ. We thus have a family of estimators called the GEL family. The most popular estimators in this class are: the Empirical Likelihood (EL) estimator which was popularized by [117], the Exponential Tilting (ET)

estimator of [100] and the Continuously Updating Estimator (CUE) of [88]. The previous ideas extend to problems of the form (2.6). [93] shows that (2.6) can be reformulated in the form of (2.8) with a number of moment equalities that diverges with n: h^{**} is the unique parameter value that satisfies for every $n \geq 1$

$$h^{**} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \sup_{\lambda \in \Lambda(h, Q_{Z,X})} \mathbb{E}_{Q_{Z,X}} \left[\psi_\gamma(\lambda' \rho(Z, h)) \otimes q_{K_n}(X) \right], \tag{2.9}$$

with $\otimes$ the Kroneker product and $q_{K_n}(\cdot)$ a vector of growing dimension K_n made of well-chosen functions. [101, 99] propose a more straightforward adaptation: they show that h^{**} verifies

$$h^{**} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \mathbb{E}_{Q_Z} \left[\sup_{\lambda \in \Lambda(h, Q_{Z|X})} \mathbb{E}_{Q_{Z|X}} \left[\psi_\gamma(\lambda' \rho(Z, h)) \mid X \right] \right], \tag{2.10}$$

where $\Lambda(h, Q_{Z|X=x}) := \bigcap_{z \in supp(Q_{Z|X=x})} \{\lambda : \psi_\gamma(\lambda' \rho(z, h)) \text{ exists}\}$. Note that even when h reduces to a finite-dimensional parameter (as is the case in [101, 99]), $\mathbb{E}_{Q_{Z|X}} [\cdot \mid X = \cdot]$ is nonparametric without further constraints. To construct GEL estimators, the previously cited articles take the sample analogue of (2.9) or (2.10) and use some nonparametric estimator to approximate $\mathbb{E}_{Q_{Z|X}} [\cdot \mid X = \cdot]$.

Very few contributions that allow h to be infinite-dimensional exist. The main ones are [116] and [40]. The class of functions $\mathcal{H}$ is always chosen as a subset of a metric space endowed with a norm $\| \cdot \|_H$. This metric space is usually taken equal to the space of square-integrable functions with respect to the Lebesgue measure or the space of uniformly bounded functions with respect to the same measure. In [116], the author focuses on models where ρ is a smooth function of h and studies the behaviour of the EL estimator based on (2.10). He uses a Nadaraya-Watson method to estimate $\mathbb{E}_{Q_{Z|X}} [\cdot \mid X = \cdot]$. His main results are the consistency of his estimator in $\| \cdot \|_H$-norm and the asymptotic normality of a specific functional of his estimator. The main limitation of this article is that unnecessary restrictions are placed on the parameter space $\mathcal{H}$ to avoid the use of regularization. In [40], the authors study the behaviour of the whole family of GEL estimators based on (2.9) in one specific model, namely the Nonparametric Quantile Instrumental Variables (NPQIV) ([41]). The NPQIV does not verify the smoothness property on the ρ function and is not covered by [116]. Larger classes of functions $\mathcal{H}$ are considered than in [116] thanks to a regularization term that is added to the estimation procedure. Their main results are the consistency with rate in $\| \cdot \|_H$-norm and the asymptotic normality of a large class of functionals of the estimator.

In Chapter 3, we study the properties of the whole GEL family of estimators for a class of ρ functions, and therefore of models, that encompasses both those studied in [116] and the NPQIV. Similar to [40], we consider a regularized estimation procedure and consider larger classes $\mathcal{H}$ than those in [116]. One specificity of our approach is that we rely on a slightly modified version of (2.10) to build our estimation method. As explained earlier, the use of regularization does not mean that $\mathcal{H}$ can be taken arbitrarily large: we assume that $\mathcal{H}$ is a subset of the space of square-integrable functions with respect to the Lebesgue measure that contains functions that are differentiable up to a certain order with all partial derivatives square-integrable. Let us denote $\| \cdot \|_{L_2(leb)}$ the norm on the space of square-integrable functions with respect to the Lebesgue measure. In our work, we prove the consistency without rate of our estimators in $\| \cdot \|_{L_2(leb)}$-norm and we derive an upper bound on the rate at which $\mathbb{E}_{Q_X} \left[\| \mathbb{E}_{Q_{Z|X}} [\rho(Z, h_n)] \|^2 \right]$ converges to 0. We prove a generic slow rate that requires weak moment assumptions and we show that the rate can be improved under more stringent moment conditions. We also discuss how those results could be used to derive consistency with rate of our estimators in $\| \cdot \|_{L_2(leb)}$-norm. As we recall in Chapter 3, to obtain the latter the key is to control the ratio $\|h - h^{**}\|_{L_2(leb)} / \mathbb{E}_{Q_X} \left[\| \mathbb{E}_{Q_{Z|X}} [\rho(Z, h)] \|^2 \right]$ uniformly over h in a suitable neighbourhood of h^{**}. This ratio measures the discrepancy between a norm in the numerator and another quantity in the denominator that can be seen loosely speaking as a weaker norm. The supremum of the ratio is sometimes called the degree of ill-posedness of the model ([37]). A large

body of work has investigated and is still actively looking for general sufficient conditions to control the degree of ill-posedness (see [39, 33] for extensive reviews). As we explain at the end of Chapter 3, we believe there is still room to find more transparent conditions to control the degree of ill-posedness. This is definitely an avenue for future research that has implications beyond the models we consider in this chapter. Other relevant extensions of our results are: i) to derive the asymptotic normality for the same class of functionals as in [40]; ii) in a more statistics-oriented way, build oracle inequalities on the estimation performance of our estimator.

Summary of Chapter 4

Even with cross-sectional data, the *i.i.d* assumption can be too restrictive. In applied econometrics, it is often plausible that the data is affected by several sources of aggregate shocks: suppose you observe several economic variables at the industry-area level. The data can be written $(Z_{i_1,i_2})_{1\leq i_1\leq n_1, 1\leq i_2\leq n_2}$, where n_1 (*resp.* n_2) is the number of industries (*resp.* areas). Observations correspond to industry-area cells and they are likely to be correlated whenever they share the same industry or area because of shocks at the industry or area level. One usually says that the data is clustered at the industry and area levels. This is an instance of multiway clustering. Polyadic data are another data type that naturally exhibit dependence: polyadic data stem from the interactions of several individuals from the same population together. Data on interactions between pairs of individuals are called dyadic for instance and are the most common. Dyadic data can be written $(Z_{i_1,i_2})_{1\leq i_1\neq i_2\leq n}$. Intuitively, polyadic data should exhibit more dependence than multiway-clustered data: in the first case, observations are dependent because of shocks that stem from a unique population while in the second case, shocks come from two distinct sources. To capture these ideas, we impose the data be jointly exchangeable in the polyadic case and separately exchangeable under multiway clustering. The two notions of exchangeability are presented in great detail in [96]. Those assumptions are powerful as they allow us to use deep and very useful probabilistic results ([89, 4, 95]) that ensure the data can be represented in terms of a series of independent shocks in the different dimensions. While separate exchangeability is a subcase of joint exchangeability, we still have to handle multiway clustering on its own: the unbalanced number of clusters in each dimension makes the problem more complicated. Quite importantly, exchangeability implies that observations remain identically distributed: the dependence we introduce is therefore very different from times series dependence.

When exchangeability is assumed instead of the *i.i.d* assumption, the construction of estimators is not affected. However, one has to show that estimators are still consistent and asymptotically normal. Existing results are mostly concerned with sample means and linear regression models: in the joint exchangeable case, asymptotic normality for sample means can be traced back to [66] and asymptotic normality for t-statistic in linear regression models is studied in [131]; under multiway clustering, [109] studies the limit in distribution of sample means when the number of relevant clustering dimensions is unknown and he shows the consistency of a bootstrap procedure (we define what a bootstrap procedure is in a few lines). A number of articles also propose estimators of the asymptotic variance for a large class of models without proving their consistency ([69, 30]). When one is interested in more models beyond the linear regression case, theoretical results for sample means are in general not enough. In the *i.i.d* case, a powerful generic approach consists in controlling the asymptotic behaviour of the empirical process associated to the model (see [137] for more details). We extend well-known results on empirical processes in the *i.i.d* case to multiway-clustered and polyadic data. To extend results, the definition of an empirical process has to be modified. As an example, under two-way clustering, the empirical process associated with the class of functions $\mathcal{F}$ is the random map $\mathbb{G}_{n_1,n_2} : f \in \mathcal{F} \mapsto$

$\frac{\sqrt{\min\{n_1,n_2\}}}{n_1 n_2}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\left(f(Z_{i_1,i_2})-\mathbb{E}_{Q_Z}[f(Z_{1,1})]\right)$. With dyadic data, the empirical process takes the form $\mathbb{G}_n : f \in \mathcal{F} \mapsto \frac{\sqrt{n}}{n(n-1)}\sum_{1\leq i_1\neq i_2\leq n}\left(f(Z_{i_1,i_2})-\mathbb{E}_{Q_Z}[f(Z_{1,2})]\right)$. The class of functions $\mathcal{F}$ depends on the model of interest. For instance, if we study the model $\mathbb{E}_{Q_Z}[\rho(Z,h)]=0 \iff h=h^{**}$, we have $\mathcal{F}:=\{\rho(\cdot,h) : h\in\mathcal{H}\}$. Observe that for every fixed $f\in\mathcal{F}$, $\mathbb{G}_{n_1,n_2}f$ and $\mathbb{G}_n f$ are asymptotically normal thanks to results on sample means. Studying the limit in distribution of the empirical process is therefore more challenging than simply requiring that the empirical process be convergent at each function f. Our main result consists in proving that empirical processes with multiway-clustered or polyadic data converge in distribution to a Gaussian process under the same assumptions as in the *i.i.d* case but the asymptotic variance formula differs from the one in the *i.i.d* setup. The Gaussian process has the following properties: this is a random function which associates to every $f\in\mathcal{F}$ a centered normal random variable with variance given by the asymptotic variance formula. This result is not directly useful to conduct inference since the asymptotic variance is unknown and has to be estimated. Instead of proposing a variance estimator, we prove the consistency of two modified version of the nonparametric bootstrap ([67]) adapted to multiway clustering and dyadic data. We explain how our bootstrap schemes are constructed with twoway clustering and dyadic data. With twoway clustering, for each dimension of clustering j we draw n_j indexes with replacement and the bootstrap process takes the form $\mathbb{G}^*_{n_1,n_2} : f\in\mathcal{F}\mapsto \frac{\sqrt{\min\{n_1,n_2\}}}{n_1 n_2}\sum_{1\leq i_1\leq n_1}\sum_{1\leq i_2\leq n_2}(V^1_{i_1}V^2_{i_2}-1)f(Z_{i_1,i_2})$, with $V^1_{i_1}$ (*resp.* $V^2_{i_2}$) the number of times the index i_1 (*resp.* i_2) is resampled. With dyadic data, we draw n indexes with replacement and the bootstrap process writes $\mathbb{G}^*_n : f\in\mathcal{F}\mapsto\frac{\sqrt{n}}{n(n-1)}\sum_{1\leq i_1\neq i_2\leq n}(V_{i_1}V_{i_2}-1)f(Z_{i_1,i_2})$. In the *i.i.d* setup, standard nonparametric bootstrap works because resampling is carried out based on independent observations. In our modified bootstrap schemes, we cannot resample at the observations' level since they are not independent at all. We thus find another level at which independence is present: the individual level that generates pairs with dyadic data and the two cluster dimensions in twoway clustering. The second main contribution of our work is the proof that the two modified bootstrap schemes are consistent asymptotically. We use our two main results to prove the asymptotic normality and validity of bootstrap-based inference for a wide class of nonlinear estimators. We also revisit the influential empirical work of [126]. The authors estimate the determinants of trade volumes between countries using explanatory variables at the country and pair-of-countries level such as a country's gross domestic product or the distance between two countries. They use a Poisson pseudo maximum likelihood (PPML) model and assume that the data is independent across pairs of countries conditional on the explanatory variables. We rerun their main specification and show that once dyadic dependence is allowed, the length of confidence intervals and p-values increase substantially.

Summary of Chapter 5

In econometrics, many parameters of interest are functions of one or several ratios of expectations and/or covariances. Leading examples are the parameters in a univariate linear regression with or without endogeneity, conditional expectations and the difference-in-difference estimand when the treatment variable is endogenous ([53]). To conduct inference on those parameters, the standard econometric approach relies on the asymptotic normality of sample means combined with the delta method (see Chapter 3 in [136]). In our work, we focus on the simple case of a ratio of expectations $\mathbb{E}_{Q_X}[X]/\mathbb{E}_{Q_Y}[Y]$ and look at the consequences of having $\mathbb{E}_{Q_Y}[Y]$ "close to zero" in terms of inference. Some deep results have already been proved on this topic: if the model does not bound $\mathbb{E}_{Q_Y}[Y]$ away from 0, it has been shown in [63] that at any confidence level an honest confidence interval in the sense of (2.2) must have infinite length with positive probability. The theorems in [63] in fact apply for any given number of observations n by dropping the limit inferior in (2.2). The question we address is also close in spirit to the

widely studied issue of weak instrumental variables: this problem shows up when the instruments have (almost) zero correlation with the endogenous variable in a linear model with endogeneity (see [8] for a recent review). The weak IV literature has proposed to build CSs that are robust to zero correlation between instruments and endogenous variables following ideas initiated in [6]. This literature has also studied the limit distribution of several estimators for the linear model under endogeneity when the correlation between endogenous variables and instruments is allowed to go to zero when the number of observations n increases ([129]). We use the latter approach to define “closeness to zero" of $\mathbb{E}_{Q_Y}[Y]$: we allow $\mathbb{E}_{Q_Y}[Y]$ to depend on n the number of observations and to go to zero as n increases. We actually allow not only $\mathbb{E}_{Q_Y}[Y]$ but also $\mathbb{V}_{Q_Y}[Y]$, $\mathbb{E}_{Q_X}[X]$ and $\mathbb{V}_{Q_X}[X]$ to depend on n and possibly go to zero. In this setting, we derive the asymptotic behaviour of the distribution of $\overline{X}_n/\overline{Y}_n - \mathbb{E}_{Q_X}[X]/\mathbb{E}_{Q_Y}[Y]$ depending on the speed at which $\mathbb{E}_{Q_X}[X]$, $\mathbb{E}_{Q_Y}[Y]$, $\mathbb{V}_{Q_X}[X]$ and $\mathbb{V}_{Q_Y}[Y]$ go to zero. We then show that when $\mathbb{E}_{Q_Y}[Y]$ (*resp.* $\mathbb{V}_{Q_Y}[Y]$) goes to zero slow enough (*resp.* fast enough), confidence intervals based on Efron’s nonparametric bootstrap are asymptotically valid in the sense of (2.1). These results are of asymptotic nature and we complement them using a completely nonasymptotic approach. To do so, we build on results from the statistics literature such as concentration inequalities (see [25] for an introduction) and the impossibility theorems of [34]. We place bounds on the second moments of $Q_{X,Y}$ as well as a lower bound on $|E_{Q_Y}[Y]|$ strictly larger than zero so that $|\mathbb{E}_{Q_X}[X]/\mathbb{E}_{Q_Y}[Y]|$ is bounded away from infinity uniformly over the model and we are not in the setup of [63]. Given theses restrictions, we show how to construct nonasymptotic confidence intervals for every confidence level below a threshold $\underline{t}_n$ with the following properties: they are almost-surely of finite length and they have the required level uniformly over the model for every finite n. The confidence intervals and $\underline{t}_n$ depend on n and the moment bounds. We further derive a confidence level $\bar{t}_n$ above which it is impossible to construct a confidence interval that contains $\overline{X}_n/\overline{Y}_n$ almost surely and that is both of required level uniformly over the model and almost surely of finite length. As a consequence, even outside the framework of [63], a large class of confidence intervals including those based on the delta method cannot be both almost surely finite and have guaranteed coverage for confidence levels too close to 1 when n is finite. We propose a criterion to appraise the reliability of the delta method in finite samples: when there exist natural upper and lower bounds on the moments of $Q_{X,Y}$, researchers can compute $\underline{t}_n$ to have an idea of the maximum confidence level (that depends on n) at which the delta method can be safely used to build confidence intervals. When no meaningful bounds can be found, we suggest a rule-of-thum criterion: simply replace these bounds by empirical moments based on the data. We advocate the use of $\underline{t}_n$ rather than $\bar{t}_n$ based on several simulation experiments. We present another impossibility result that gives a minimal confidence interval’s length below which the said confidence interval cannot have uniform coverage. We illustrate our asymptotic and nonasymptotic findings on a simple application to gender wage disparities using French administrative data.

Summary of Chapter 6

This chapter proposes a Stata package implementing the different statistical tools introduced in [53]. [53] starts from Rubin’s causal framework and assume $Z_o = DZ_o(1) + (1-D)Z_o(0)$. Each individual is also attributed a random pair (G,T): T is the random time period or cohort of that individual and G indicates whether that individual belongs to a group with a stable or increasing intensity of treatment between time periods/cohorts. G identifies treatment ($G=1$) and control groups ($G=0$) in [53]. Letting S identify all the individuals in the treatment group that would switch from non-treatment to treatment should they be observed at different time periods/cohorts, [53] gives non-nested sets of assumptions that allow to recover the quantity $\Delta = \mathbb{E}_{Q_{(Z_o(1),Z_o(0))|S,T=1}}[Z_o(1) - Z_o(0) \mid S, T=1]$ with three different

estimands. The parameter Δ is called a local average treatment effect (LATE) and was introduced in [92]. One of the estimands denoted W_{DID} is not new and is widely used in practice while the other two coined W_{TC} and W_{CIC} are new. Under the identification conditions that motivate the W_{CIC} estimand, the authors prove the stronger fact that $Q_{Z_o(1)|S,T=1}$ and $Q_{Z_o(0)|S,T=1}$ are identified as well as local quantile treatment effects (LQTEs) $\tau_\delta = q_{Q_{Z_o(1)|S,T=1}}(\delta) - q_{Q_{Z_o(0)|S,T=1}}(\delta)$. On top of these identification results, [53] proposes estimators for the four estimands and show their asymptotic normality. One major takeaway of [53] is to show that the conditions required to identify the popular W_{DID} estimand may be implausible in certain settings in which case W_{TC} and W_{CIC} can be useful alternatives. Our contribution is to make the proposed estimation procedures available on the statistical software Stata that is widely used in applied econometrics. On top of computing the estimators, we build in the package 95% confidence intervals on Δ and τ_δ based on the bootstrap as well as statistical tests to see whether the estimands of Δ are significantly different. Inference can be made robust to one-way clustering. In [53] and [54], several extensions are considered: analogues of Δ are defined with multiple time periods/cohorts, treatment levels and groups and their identification is proved under adapted conditions; the results are also extended to setups where assumptions are valid conditional on additional covariates Z_e; corresponding estimators are proposed. When additional covariates are included, estimators of W_{DID}, W_{TC} and W_{CIC} require to estimate quantities of the form $\mathbb{E}_{Q_{Z_o|G,T,X}}[Z_o \mid G,T,X]$ and $\mathbb{E}_{Q_{D|G,T,X}}[D \mid G,T,X]$. [54] shows the asymptotic normality of estimators of W_{DID}, W_{TC} and W_{CIC} when conditional expectations are estimated nonparametrically using polynomial regressions. Our Stata package supports these extensions as well. The conditional expectations $\mathbb{E}_{Q_{Z_o|G,T,X}}[Z_o \mid G,T,X]$ and $\mathbb{E}_{Q_{D|G,T,X}}[D \mid G,T,X]$ can be estimated by ordinary least squares, Probit or Logit (when Z_o or D is binary) or polynomial nonparametric regression. The order of the polynomial regression can be specified by the user or automatically chosen via 5-fold cross-validation based on a mean squared error criterion (see [135] for definitions). Similar to [54], we revisit the empirical work of [76] to show how to use our Stata command and to emphasize differences that can be found when using the W_{DID} estimand rather than the W_{TC} one for instance. We conclude with a substantial simulation study to check the performance of our estimators in moderately large samples. For the data generating process that we select, we run 1,000 replications of it for each of three different sample sizes, namely 400, 800 and 1,600. We assess the quality of our estimators based on average bias, average mean squared error and coverage rate, where the average and coverage rate are computed with the 1,000 replications.

Chapter 3

Nonparametric estimation in conditional moment restricted models via Generalized Empirical Likelihood

Abstract

In this paper we address the issue of estimating a functional parameter h_0 identified by a set of conditional moment restrictions. In particular the arguments of h_0 are allowed to be endogenous, a situation we refer to as nonparametric endogeneity. The models we consider can be written as inverse problems of the form $\|Th\| = 0 \iff h = h_0$ for some (nonlinear) integral operator T with $\|\cdot\|$ the norm on the codomain of T. What is more T is unknown and has to be estimated. To recover h_0, we propose an estimator $\widehat{h}_n$ based on a penalized kernel Generalized Empirical Likelihood (GEL) procedure. For a class of models that encompasses both the Nonparametric Instrumental Variables mean (NPIV) and quantile (NPQIV) regressions, we derive the consistency of our estimator in $L_2(P)$ norm where P is the unknown distribution of the data. We also obtain an upper bound on the rate of decrease of $\|T\widehat{h}_n\|$ to 0. We discuss how this last result can be used to control the convergence rate of $\widehat{h}_n$ in $L_2(P)$ norm. Our results notably complement [116] and [40]: the former propose a GEL estimator for a class of models that includes the NPIV but not the NPQIV while the latter focus on the NPQIV only.

Keywords: NPIV, NPQIV, penalized nonparametric regression, statistical inverse problems.
Based on [85] : Guyonvarch Y., Nonparametric estimation in conditional moment restricted models via Generalized Empirical Likelihood.

3.1 Introduction

Our goal in this article is to estimate infinite-dimensional parameters in models subject to endogeneity, a situation we refer to as “nonparametric endogeneity”. More specifically, we want to study the performance of the family of Generalized Empirical Likelihood (GEL) estimators in this context.

The issue of nonparametric endogeneity in structural estimation has received growing attention over the past 15 years. The challenges put forward by this question are well-exemplified by the classical

problem of estimating budget share Engel curves.[1] Budget share Engel curves capture how budget shares devoted to a specific good vary with total consumption expenditures of households for a given system of relative prices, potentially controlling for additional households' characteristics. It has been well documented ([21, 23, 22]) that: (i) for consumer maximization theory to hold, total expenditure and additional households' characteristics can enter linearly or additively in the regression function only if strong restrictions are imposed on the utility function; (ii) total expenditure is likely to be endogenous.

Point (i) illustrates that reduced form estimation may come at odds with economic theory when it does not allow for enough flexibility. Point (ii) underlines the general fact that accounting for endogeneity is often a key ingredient when estimating structural economic relations. In that respect, nonparametric econometric tools that allow for the presence of endogeneity are crucial.

Continuing with the Engel curves example, let Z_o stand for the budget share spent on food for instance, Z_h stand for the log of total consumption expenditures (*a priori* endogenous) and X be a vector of excluded instruments. We omit the presence of additional household characteristics for simplicity. We further let P stand for the distribution of (Z_o, Z_h, X) and P^V be the marginal distribution of any subset V of (Z_o, Z_h, X). If one is interested in studying the average impact of Z_h on Z_o, the canonical nonparametric instrumental mean regression model (NPIV) writes

$$\mathbb{E}\left[Z_o - h_0(Z_h) \mid X\right] = 0 \quad P^X - \textit{almost surely} \quad (P^X - a.s).$$

If one is willing to recover the effect of Z_h on the τ-th quantile of the (conditional) distribution of Z_o, we are left with the nonparametric instrumental quantile regression model (NPQIV)

$$\mathbb{E}\left[\mathbb{1}\left\{Z_o \leq h_0(Z_h)\right\} - \tau \mid X\right] = 0 \quad P^X - a.s.$$

The NPIV and NPQIV are in fact two instances of a broader class of models we are interested in and that we now introduce. Let $(Z_o^t, Z_h^t, X^t)^t$ be a random vector with distribution P and support $\mathcal{Z}_o \times \mathbb{R}^{d_{z_h}} \times [0,1]^{d_x}$ where $\mathcal{Z}_o \subseteq \mathbb{R}^{d_{z_o}}$. We denote $(Z_o^t, Z_h^t)^t$ by Z.[2] Define the parameter space $\mathcal{H} := \{h : \mathcal{Z}_h \to \mathbb{R} : \|h\|_2 < \infty\} \subseteq L_2(P)$ where $L_2(P)$ is the space of square integrable functions of Z_h with respect to P and its norm is denoted by $\|\cdot\|_2$. The true parameter h_0 is assumed to belong to $\mathcal{H}$ and is characterized via the following restriction

$$\mathbb{E}[\rho(Z,h)|X] = 0 \quad P^X - a.s \quad \iff \quad h = h_0, \tag{3.1}$$

where "$h = h_0$" means $\|h - h_0\|_2$ is null, $\rho : \mathbb{R}^{d_{z_o}+d_{z_h}} \times \mathcal{H} \to \mathbb{R}^d$ is a vector of known functions and $\mathbb{E}[\cdot]$ denotes the expectation with respect to P. In the following we denote $\mathbb{E}[\rho(Z,h)|X = x]$ by $m(x,h)$. We can verify that the NPIV model corresponds to the choice $\rho(z,h) = z_o - h(z_h)$ while the NPQIV model is obtained by setting $\rho(z,h) = \mathbb{1}\{z_o \leq h(z_h)\} - \tau$. Our goal is to estimate h_0 consistently in $\|\cdot\|_2$-norm. Providing primitive conditions that ensure existence and unicity of h_0 is not trivial at all and lies beyond the scope of this article. [61] and [7] provide exhaustive discussions on the topic.

The difficulty of estimating h_0 in (3.1) is now well-understood: as explained in [32], the operator $T : h \mapsto m(X,h)$ that maps $L_2(P^{Z_h})$ into $L_2(P^X)$ is not continuously invertible in general when $\mathcal{H}$ is not a compact subset of $L_2(P)$. T is also unknown here since P is not specified and has to be estimated. Consequently, estimating T consistently and inverting it is not enough in general to estimate h_0 consistently in $\|\cdot\|_2$-norm when $\mathcal{H}$ is not a compact subset of $L_2(P)$. Estimating h_0 consistently in

[1] See *e.g* [2] for other interesting examples such as estimation of production functions or multiple-period choice models.

[2] Note that we restrict ourselves to cases where Z and X do not have elements in common so that we do not allow for exogenous regressors in the NPIV or NPQIV models. Allowing for some overlap between X and Z would increase the technicality of several steps in the proofs. We leave this task for future research.

this framework is called an ill-posed statistical inverse problem. For clear and concise introductions to ill-posed statistical inverse problems with many motivating examples, we refer the reader to [32] and [35].

To have a chance to recover h_0, we need to *regularize* T when inverting it. We start by presenting a method which does not lead to GEL-type estimators but that is a very natural way to tackle this problem. We observe that for every strictly positive weight function $w(\cdot)$ independent of h, (3.1) is equivalent to $h_0 = \text{argmin}_{h\in\mathcal{H}} \mathbb{E}\left[\|m(X,h)\|^2 w(X)\right]$. Intuitively, solving $\min_{h\in\mathcal{H}_n} \mathbb{E}\left[\|m(X,h)\|^2 w(X)\right] + \alpha_n Pen(h)$ where $\mathcal{H}_n$ grows dense in $\mathcal{H}$ with n and $\alpha_n \to 0$ may help if some restrictions are placed on $\mathcal{H}_n$, α_n and $Pen(\cdot)$. Replacing $\mathbb{E}\left[\|m(X,h)\|^2 w(X)\right]$ with an estimator based on a sample $(Z_i, X_i)_{i=1}^n \overset{i.i.d}{\sim} P$ yields the empirical regularized version of (3.1):

$$\widehat{h}_n \in \underset{h\in\mathcal{H}_n}{\text{argmin}} \frac{1}{n}\sum_{i=1}^{n} \|\widehat{m}(X_i,h)\|^2 \widehat{w}(X_i) + \alpha_n Pen(h). \tag{3.2}$$

We call (3.2) a Generalized-Method-of-Moments (GMM) approach. A vast literature has studied the performance of $\widehat{h}_n$ defined in (3.2) to recover h_0 for different choices of $\mathcal{H}_n$, $\widehat{m}(\cdot)$, $\widehat{w}(\cdot)$ and $Pen(\cdot)$. [86], [48], [36], [44], [91] and [13] focused on the NPIV, [41] and [90] on the NPQIV, [3], [112], [20], [37] and [64] studied classes of models satisfying (3.1).

It is important to note that the choice of $Pen(\cdot)$ may put additional restrictions on the model $\mathcal{H}$ to which h_0 belongs. Let leb stand for the Lebesgue measure. Choosing *e.g* the squared $L_2(leb)$-norm of the m-th derivative of h is a classical choice which imposes that h_0 is at least m times differentiable (otherwise the penalty is not informative about h_0). As explained below we will make this smoothness assumption on h_0. Imposing shape constraints on h_0 on top of smoothness ones is a method that is gaining popularity due to its natural connection with economic theory and its finite-sample performances ([44], [91]). We do not investigate here the advantages of imposing additional shape restrictions on the function to retrieve but we acknowledge it could be a promising line of research for future work.

We focus on the GEL family of estimators which offers an alternative to (3.2). The construction of those estimators is less straightforward than for GMMs and we postpone a full description of the method to Section 3.2. In the simpler case where h is replaced with a finite-dimensional parameter θ, GEL-type estimators have been studied extensively ([100], [93], [101], [119]). They have also been shown to exhibit nice theoretical properties ([113]). However the properties of these estimators are much less known for models that display nonparametric endogeneity. [116] proposed an early contribution using kernel-regression techniques for a class of models that nests the NPIV but not the NPQIV, under the assumption that $\mathcal{H}$ is a compact subset of $L_2(P)$. In a very recent contribution, [40] prove the consistency with rate in $\|\cdot\|_{L_2(leb)}$-norm in the NPQIV model without a compactness assumption on $\mathcal{H}$ for estimators based on sieve estimation techniques. They also show the asymptotic normality of an estimator of $\theta_0 := \mathbb{E}[\mu(Z_h)\partial_j h_0(Z_h)]$ for some weight $\mu(\cdot)$ based on estimating h_0.[3] They finally discuss the impact of the degree of ill-posedness of (3.1) (*i.e* how difficult it is to regularize the problem) on the semi-parametric efficiency bound in estimating θ_0. Our goal is to provide a unifying framework to prove consistency in both the NPIV and NPQIV and more generally for all models that satisfy (3.1) using a kernel-regression approach similar to [116].

[40] emphasize that the parameter they are mainly interested in is θ_0. As a matter of fact, parameters of interest for economists are often *functionals* of h_0 rather than h_0 itself. This is nicely presented in [49] in the NPIV setting: the author gives the example of the average marginal effect with respect to the j-th argument of Z_h: $\theta_0 = \mathbb{E}\left[\partial_j h(Z_h)\right]$. Assuming Z_h is continuously distributed with Lebesgue density f_{Z_h}, he also mentions the expectation of Z_o under a counterfactual distribution $f^c_{Z_h}$, $\theta_0 = \mathbb{E}\left[\frac{f^c_{Z_h}(Z_h)}{f_{Z_h}(Z_h)} h_0(Z_h)\right]$.

[3] $\partial_j h_0$ stands for the partial derivative with respect to the j-th argument of h_0.

[38] provide additional examples of functionals of h_0 that are of interest in economics, notably to test whether h_0 is linear or not. They further derive asymptotic normality results when estimating a functional of h_0 in a large class of models that are subject to (3.1). Their results rely on a GMM approach.

While plugging-in a consistent estimator of h_0 is not the only way to estimate θ_0 consistently ([125]), this remains a natural approach. In this article we focus on the first step, namely estimating h_0. Unlike [40], we do not derive the consistency and asymptotic normality of functionals of h_0 based on estimating h_0 with GELs. The next big step for us is therefore to take advantage of the consistency results we derive for GELs to prove the asymptotic normality of functionals of h_0 not only in the NPQIV but also in all models that satisfy (3.1).

In Section 3.2 we introduce the theoretical background that justifies the use of GELs to estimate h_0 in (3.1) and we detail the construction of our estimation procedure. In Section 3.3 we present and discuss the assumptions that we impose, we prove the consistency of our estimator $\widehat{h}_n^{GEL}$ in $\|\cdot\|_2$-norm and we give an explicit rate at which $\mathbb{E}\left[\|m(X,\widehat{h}_n^{GEL})\|^2\right]$ goes to zero. We also explain how this last result can be used to derive an explicit consistency rate for $\widehat{h}_n^{GEL}$ in $\|\cdot\|_2$-norm. Section 5.7 concludes. The results from Section 3.3 are proved in Section 3.5. Section 3.6 gathers all the lemmas and their proofs.

Notation. We denote by $\mathbb{P}$ the probability taken with respect to $P^{\otimes n}$. For any probability measure μ, we denote $supp(\mu)$ its support. We use a.s (*resp.* w.p.a.1) to denote with probability 1 (*resp.* with probability approaching one). For two measures μ_1 and μ_2, $\mu_1 \ll \mu_2$ means that μ_1 is absolutely continuous with respect to μ_2. For all $x \in \mathbb{R}^p$, we let $(x^{(2.)})_{l=1}^p$ denote its p components. For all $(\beta, x) \in \mathbb{N}^p \times \mathbb{R}^p$ and a function f of x, let $|\beta| = \beta_1 + \cdots + \beta_p$, $x^\beta = (x^{(1)})^{\beta_1} \times \cdots \times (x^{(m)})^{\beta_m}$ and $\nabla^\beta f = \frac{\partial^{|\beta|} f}{\partial (x^{(1)})^{\beta_1} ... \partial (x^{(m)})^{\beta_m}}$. We denote by $\|\cdot\|$ the Euclidean norm and the matrix norm induced by the Euclidean norm and by $\|\cdot\|_2$ the norm in $L_2(P)$. For every $p \in [1, +\infty]$ and for functions with domain $\mathcal{Z}_h$ that are m times differentiable, we let $\|h\|_{m,p}$ stand for $\left(\sum_{0\leq l\leq m} \left\|\nabla^l h\right\|_{L^p(leb)}^p\right)^{1/p}$ (when $p = +\infty$, we simply have a sup-norm). For a square matrix A, $\sigma_{min}(A)$ stands for the smallest eigenvalue of A. For every $l \in \{1, ..., d\}$, $m_l(x, h) = \mathbb{E}\left[\rho_l(Z, h) \mid X = x\right]$. Let $\mathcal{H}$ be a class of functions mapping $\mathcal{Z}_h$ to $\mathbb{R}$ and $\mathcal{H} \subseteq \mathcal{S}$ where $(\mathcal{S}, q)$ is a metric space. $H \in \mathcal{S}$ is called an envelope for $\mathcal{H}$ if for every $z_h \in \mathcal{Z}_h$, $\sup_{h\in\mathcal{H}} |h(z_h)| \leq H(z_h)$. For every $\epsilon > 0$, the q-bracketing number $N_{[\,]}(\epsilon, \mathcal{H}, q)$ is (when it exists) the smallest number m of pairs of functions in $(\mathcal{S}, q)$, $\{(l_i, u_i)\}_{i=1}^m$, such that for every $i \in \{1, ..., m\}$, $l_i \leq u_i$, $q(u_i, l_i) \leq \epsilon$ and for every $h \in (\mathcal{H}, q)$, there exists $i \in \{1, ..., m\}$ that satisfies $l_i \leq h \leq u_i$. For every $\epsilon > 0$, the q-covering number $N(\epsilon, \mathcal{H}, q)$ is (when it exists) the smallest number of closed balls of radius ϵ with centers in $\mathcal{H}$ needed to cover $(\mathcal{H}, q)$. For every $\gamma \in \mathbb{R}$ and $h : z_h \mapsto h(z_h)$, $||h||_{\infty,\gamma} := \sup_{z_h\in\mathcal{Z}_h} |h(z_h)\langle z_h\rangle^{-\gamma}|$ with $\langle z_h\rangle = \left(1 + ||z_h||^2\right)^{1/2}$. For every $M_0 > 0$, let $\mathcal{H}^{M_0} := \{h \in \mathcal{H} : Pen(h) \leq M_0\}$ with $Pen(h)$ some positive functional defined later. We sometimes use $a \vee b$ (*resp.* $a \wedge b$) instead of $\max\{a, b\}$ (*resp.* $\min\{a, b\}$). $\lceil\cdot\rceil$ is the ceiling function, *i.e* $\lceil x\rceil$ is the smallest integer larger than or equal to x.

3.2 A general presentation of GEL estimators

In this section, we show that the problem in (3.1) can be rewritten as a constrained minimization problem over sets of probability measures. We denote this reformulation as a Generalized Minimum Contrast (GMC) version of (3.1). There are actually a collection of GMCs depending on how the distance between probability measures is computed. We then show that each GMC problem admits a dual expression which leads to the population counterpart of a GEL criterion. Finally we present the empirical version of each GEL procedure. This presentation is an extension of Section 3 in [99] to conditional moment

restrictions and infinite-dimensional parameter spaces.

3.2.1 From GMCs to GELs

Let

$$\mathcal{D}_\phi\left(Q \mid P^{Z|X=x}\right) := \begin{cases} \int_{\mathcal{Z}} \phi\left(\frac{dQ}{dP^{Z|X=x}}\right) dP^{Z|X=x} \text{ if } Q \ll P^{Z|X=x} \\ +\infty \text{ otherwise} \end{cases}$$

where ϕ is a *discrepancy*, *i.e* a convex function such that $\mathcal{D}_\phi\left(. \mid P^{Z|X=x}\right)$ is uniquely minimized at $P^{Z|X=x}$.

Let $\mathcal{Q}(h) := \left\{Q \in \mathcal{M} : \int \rho(z,h) dQ = 0\right\}$, where $\mathcal{M}$ is the set of all probability measures on $\mathcal{Z}$. When h_0 satisfies (3.1), it is the unique minimizer of the following optimization problem for every x in a set of measure 1 under P^X

$$\inf_{h\in\mathcal{H}} \inf_{Q\in\mathcal{Q}(h)} D_\phi\left(Q \mid P^{Z|X=x}\right). \tag{3.3}$$

This optimization problem is called the primal GMC problem. It is the first step towards constructing GEL estimators. Note though that this primal problem is not directly useful to build an estimation procedure.

First, even though h_0 minimizes the problem for every x in a set of P^X- measure 1, the empirical counterpart of (3.3) evaluated at different x's would yield different minimizers. There is *a priori* no clear rule to choose between different empirical minimizers. A solution is to turn the identifying equation (3.1) into a continuum of unconditional moment restrictions (*cf.* [93], [31], [40]). Since a continuum of moment restrictions cannot be handled in practice, one has to consider a finite but growing number of unconditional moment restrictions, which introduces a regularization bias that does not arise when one directly works with the conditional moment restriction given in (3.1).

The second issue is that for every x, the problem amounts to solving two nested infinite-dimensional minimization problems. This problem can be addressed since the program (3.3) is a convex constrained program. It is indeed an established fact in convex functional analysis ([24]) that for every x in a set of measure 1 under P^X, h_0 is the solution of the so-called Minimum Contrast dual optimization problem

$$h_0 = \underset{h\in\mathcal{H}}{\operatorname{argmin}} \sup_{(\lambda_1,\lambda_2)\in\Lambda_{1,2}(h,P^{Z|X=x})} \left\{\lambda_1 - \int \phi^*(\lambda_1 + \lambda_2^t \psi(z,h)) dP^{Z|X=x}\right\} \tag{3.4}$$

where ϕ^* is the convex conjugate of ϕ ([26]) and

$$\Lambda_{1,2}(h, P^{Z|X=x}) := \bigcap_{z\in supp(P^{Z|X=x})} \left\{(\lambda_1,\lambda_2) \in \mathbb{R}^{d+1} : \phi^*(\lambda_1 + \lambda_2^t\psi(z,h)) \text{ exists }\right\}.$$

We can see that for every x, the infinite-dimensional minimization over conditional probability distributions is replaced with a minimization over a subset of $\mathbb{R}^{d+1}$.

We now restrict ourselves to the *Cressie-Read* family of discrepancies, for which the dual problem (3.4) can be further simplified. In the *Cressie-Read* family, discrepancies are indexed by a parameter γ and for every γ, ϕ_γ takes the form

$$\phi_\gamma(u) = \frac{2}{\gamma \times (\gamma+1)}(u^{-\gamma} - 1).$$

For any ϕ_γ, following [113] and [99] we can write (3.4) as

$$\begin{aligned} h_0 &= \underset{h\in\mathcal{H}}{\operatorname{argmin}} \sup_{\lambda\in\Lambda(h,P^{Z|X=x})\subseteq\mathbb{R}^d} \int \psi_{\phi_\gamma}(\lambda^t\rho(z,h)) dP^{Z|X=x} \\ \iff h_0 &= \underset{h\in\mathcal{H}}{\operatorname{argmin}} \sup_{\lambda\in\Lambda(h,P^{Z|X=x})\subseteq\mathbb{R}^d} \mathbb{E}\left[\psi_{\phi_\gamma}(\lambda^t\rho(Z,h)) \mid X=x\right], \end{aligned} \tag{3.5}$$

where $\Lambda(h, P^{Z|X=x}) := \bigcap_{z \in supp(P^{Z|X=x})} \{\lambda \in \mathbb{R}^d : \psi_{\phi_\gamma}(\lambda^t \rho(z,h)) \text{ exists } \}$ and ψ_{ϕ_γ} is the GEL criterion associated with ϕ_γ. ψ_{ϕ_γ} satisfies

$$\psi_{\phi_\gamma}(u) = \frac{2}{\gamma}\left[-(\gamma+1)\frac{u+1}{2}\right]^{\frac{\gamma}{\gamma+1}} - \frac{2}{\gamma(\gamma+1)}.$$

As detailed in [113], those GEL criteria are concave functions defined on an open interval $\mathcal{V}_\gamma$ containing 0. This implies that for every x and h, the set $\Lambda(h, P^{Z|X=x})$ contains $\mathbf{0}_d$, the null element in $\mathbb{R}^d$. For $\gamma = 0$ or $\gamma = -1$, it can be checked that ψ_{ϕ_γ} is well-defined as well. For every $\gamma \in \mathbb{R}$, ψ_{ϕ_γ} is twice-continuously differentiable on $\mathcal{V}_\gamma$ with Lipschitz second derivative on any compact subinterval of $\mathcal{V}_\gamma$ that contains strictly 0. In [113], the authors argue that it is relevant to focus on GEL criteria that satisfy $\psi'_{\phi_\gamma}(0) \neq 0$ and $\psi''_{\phi_\gamma}(0) < 0$. As a matter of fact, the most popular GEL criteria satisfy this constraint: when $\gamma = 0$, we obtain the Empirical Likelihood (EL) criterion with $\psi_{\phi_0}(u) = \log(1+u)$ ([117]); when $\gamma = -1$, we have the Exponential Tilting (ET) function $\psi_{\phi_{-1}}(u) = -e^u$ ([100]); when $\gamma = 1$, we get the Continuous Updating Estimator (CUE) function $\psi_{\phi_1}(\cdot)$ which is quadratic in u ([88]). For GEL criteria that satisfy $\psi'_{\phi_\gamma}(0) \neq 0$ and $\psi''_{\phi_\gamma}(0) < 0$, it is without loss of generality to assume that up to a renormalization $\psi'_{\phi_\gamma}(0) = \psi''_{\phi_\gamma}(0) = -1$. In the sequel, we only focus on GEL criteria that satisfy $\psi'_{\phi_\gamma}(0) = \psi''_{\phi_\gamma}(0) = -1$.

We drop the dependence of ψ_{ϕ_γ} on ϕ_γ for notational convenience. The expression in (3.5) is still not very convenient from an estimation perspective: in finite samples, the empirical counterpart of (3.5) would likely give different solutions when evaluated at different x values. It is actually possible to get round that issue. We remark that for every $h \in \mathcal{H}$

$$\mathbb{E}\left[\sup_{\lambda \in \Lambda(h, P^{Z|X})} \mathbb{E}\left[\psi(\lambda^t \rho(Z,h)) \mid X\right]\right] \geq \mathbb{E}\left[\inf_{h \in \mathcal{H}} \sup_{\lambda \in \Lambda(h, P^{Z|X})} \mathbb{E}\left[\psi(\lambda^t \rho(Z,h)) \mid X\right]\right],$$

with equality only at h_0. Combining this with (3.5) and the fact that $h_0 \in \mathcal{H}$ allows us to rewrite (3.5) as

$$h_0 = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, \mathbb{E}\left[\sup_{\lambda \in \Lambda(h, P^{Z|X})} \mathbb{E}\left[\psi(\lambda^t \rho(Z,h)) \mid X\right]\right].$$

This new way to define h_0 is a version of (3.5) integrated over the distribution P^X. With this new expression, we can clearly see the link with the other existing GEL approach which turns the initial problem (3.1) into a continuum of unconditional moment restrictions.

A final modification of the GEL procedure is in order before building its empirical counterpart. Let $w(\cdot)$ be a weight function that satisfies $w(X) > 0$ P^X-a.s. As $w(X)$ does not depend on λ and h, we can see that h_0 is uniquely determined as follows

$$h_0 = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, \mathbb{E}\left[\sup_{\lambda \in \Lambda(h, P^{Z|X})} \mathbb{E}\left[\psi(\lambda^t \rho(Z,h)) \mid X\right] w(X)\right]. \tag{3.6}$$

As explained below, this last trick is only useful from an estimation point of view.

3.2.2 Construction of the estimation procedure

We assume from now on that P^X has a Lebesgue density f_X. By definition $f_X(X) > 0$ P^X-a.s and $f_X(X)$ does not depend on λ and h so that we can take $w(\cdot) = f_X(\cdot)$ in (3.6). In (3.6) we want to estimate: (i) the outside expectation with respect to P^X as an average over the empirical distribution $\frac{1}{n}\sum_{i=1}^n \delta_{\{X_i\}}$; (ii) the quantity $\mathbb{E}\left[\psi(\lambda^t \rho(Z,h)) \mid X = \cdot\right] f_X(\cdot)$ with a Nadaraya-Watson approach. Let $\widehat{f}_X(\cdot)$ be a Nadaraya-Watson estimator of $f_X(\cdot)$. Multiplying $\mathbb{E}\left[\psi(\lambda^t \rho(Z,h)) \mid X = \cdot\right]$ by $f_X(\cdot)$ avoids handling the quantity $1/\widehat{f}_X(\cdot)$ which appears in the kernel estimator of $\mathbb{E}\left[\psi(\lambda^t \rho(Z,h)) \mid X = \cdot\right]$ ([135]). It is

appealing in practice since $1/\widehat{f}_X(\cdot)$ can be quite unstable, in particular close to the boundary of the support of P^X. This trick has been employed in many articles in which estimating $\mathbb{E}\left[\psi(\lambda^t\rho(Z,h)) \mid X=\cdot\right] f_X(\cdot)$ suffices ([90], [103], [13]).

Let $\Lambda_n(h) := \left\{\lambda \in \mathbb{R}^d : \psi\left(\lambda'\rho(Z_j,h)\right) \text{ well-defined } \forall j \in \{1,...,n\}\right\}$. Given the discussion in the previous paragraph, for every $h \in \mathcal{H}$ we estimate
$\mathbb{E}\left[\sup_{\lambda\in\Lambda(h,P^{Z|X})} \mathbb{E}\left[\psi(\lambda^t\rho(Z,h)) \mid X\right] f_X(X)\right]$ as

$$\frac{1}{n}\sum_{i=1}^{n} \sup_{\lambda\in\Lambda_n(h)} \frac{1}{nb_n^{d_x}} \sum_{j=1}^{n} K(X_i,X_j,b_n)\psi(\lambda^t\rho(Z_j,h)), \tag{3.7}$$

where $K(\cdot,\cdot,\cdot)$ is a nonparametric kernel function specified later and b_n is the bandwidth parameter that is strictly positive and decreases to 0 as n goes to infinity. Subsequently K_{ij} is a shortcut for $K(X_i,X_j,b_n)$.

As mentioned in the introduction, in order to regularize the estimation of h_0, we can restrict $\mathcal{H}$ to be a subspace of $L^2(P)$ of m-times differentiable functions. We take $\mathcal{H}$ equal to

$$\mathcal{H} := \left\{h \in L^2(P) : \|h\|_{m,2} < +\infty\right\}.$$

The choice $p=2$ corresponds to choosing $\mathcal{H}$ as a L^2 Sobolev space of smoothness m. Another popular choice corresponds to $p=+\infty$ and yields so-called Hölder spaces. Our analysis would go through with only minor changes if we picked $p=+\infty$. Remark that Sobolev and Hölder spaces are well-defined for non-integer degrees of smoothness m but the definition of the norm $\|\cdot\|_{m,p}$ has to be adapted ([114]). The results we will present in subsequent sections apply to cases where $m \in \mathbb{R}_+ \setminus \mathbb{N}$ as well. The space $\mathcal{H}$ we choose consists of bounded functions. We do not allow for unbounded functions (the weighted Sobolev/Hölder case) to keep the exposition simple.

Another crucial remark is in order: the restriction we place on $\mathcal{H}$ is only useful to regularize the estimation problem if there exists some constant $C>0$ such that for every $h \in \mathcal{H}$: $\|h\|_2 \leq C\|h\|_{2,m}$. This is verified whenever P^{Z_h} admits a bounded Lebesgue density for instance.

When $\|h\|_2 \leq C\|h\|_{2,m}$ for every $h \in \mathcal{H}$, we can combine (3.6) and (3.7) and add a penalty term $Pen(h) = \|h\|_{2,m}^2$ to the objective function to define h_n^* as any element in $\mathcal{H}$ that satisfies

$$\widehat{\mathcal{L}}_n(h_n^*) + \alpha_n Pen(h_n^*) \leq \inf_{h\in\mathcal{H}} \left\{\widehat{\mathcal{L}}_n(h) + \alpha_n Pen(h)\right\} + R_n,$$

where $\widehat{\mathcal{L}}_n(h) := \frac{1}{n}\sum_{i=1}^n \sup_{\lambda\in\Lambda_n(h)} \frac{1}{nb_n^{d_x}} \sum_{j=1}^n K_{ij}\psi(\lambda^t\rho(Z_j,h))$, $R_n = O_P(n^{-1})$ and α_n is a second tuning parameter that goes to 0 as n goes to infinity. We need to define h_n^* as an *approximate* minimizer as we allow for functions $\rho(\cdot,\cdot)$ that are nonsmooth in h ([111], [119]). We impose R_n goes to zero fast enough which ensures that this term does not have an impact on the theoretical analysis.

Unfortunately, h_n^* is infeasible in practice: it is computationally impossible to optimize an objective over an infinite-dimensional space. A solution is to replace $\mathcal{H}$ with a sequence of finite-dimensional spaces $\{\mathcal{H}_n\}_{n\geq 1}$ with the following property: for every $h \in \mathcal{H}$, there exists a sequence $(h_n)_{n\geq 1}$ such that for every $n \geq 1$ $h_n \in \mathcal{H}_n$ and $\|h-h_n\|_{L_2(leb)} = o(1)$. Such a sequence of spaces $(\mathcal{H}_n)_{n\geq 1}$ is said to grow dense in $\mathcal{H}$ as n goes to infinity and is called a sequence of finite-dimensional *sieve spaces*. We choose

$$\mathcal{H}_n := \left\{h(\cdot) = \sum_{k=1}^{\varphi(b_n^{-1})} b_k q_k(\cdot) : (b_1,...,b_{\varphi(b_n)})^t \in \mathbb{R}^{\varphi(b_n)}\right\},$$

where $\varphi(\cdot)$ is integer-valued, increasing, $\varphi(1) \geq 1$, $\lim_{u\to+\infty}\varphi(u) = +\infty$ and $(q_k(\cdot))_{k\geq 1}$ is a known family of functions that are square-integrable over $\mathcal{Z}_h$ with respect to the Lebesgue measure. $(q_k(\cdot))_{k\geq 1}$ can be chosen as an orthonormal basis of $L_2(leb)$ but this is not necessary. We only introduce sieve

spaces for a computational reason. Our theoretical results apply directly to h_n^*. This implies that from a theoretical point of view, the regularization power of the sieve dimension $\varphi(b_n^{-1})$ does not really play a role here. In particular regularization cannot be obtained without the use of a penalty function $Pen(\cdot)$ except if additional restrictions are placed on $\mathcal{H}$.[4] In practice this leads to choosing $\varphi(b_n^{-1})$ as large as possible subject to numerical tractability. Finally note that the spaces $\mathcal{H}_n$ we use have two other very nice computational properties: they are linear and unconstrained.

We eventually define our estimator $\widehat{h}_n$ as any element in $\mathcal{H}_n$ which satisfies

$$\widehat{\mathcal{L}}_n(\widehat{h}_n) + \alpha_n Pen(\widehat{h}_n) \leq \inf_{h \in \mathcal{H}_n} \left\{ \widehat{\mathcal{L}}_n(h) + \alpha_n Pen(h) \right\} + R_n. \tag{3.8}$$

The penalty functional $Pen(\cdot)$ is known here since it depends on the $\|\cdot\|_{L_2(leb)}$-norm. However this penalty must still be approximated in practice (only a discretized version of it can be computed numerically). We do not introduce this additional difficulty in the analysis for the sake of simplicity.

3.3 Results

3.3.1 Consistency

We present the set of assumptions that we use to prove the consistency of $\widehat{h}_n$ in $\|\cdot\|_2$-norm. We notably compare those assumptions with what would be needed to ensure consistency of $\widehat{h}_n$ in $\|\cdot\|_2$-norm with the GMM approach recalled in (3.2).

Assumption 3.1. *$\{Z_i, X_i\}_{i=1}^n$ are i.i.d. copies of $(Z, X) \sim P$. P has support $\mathcal{Z}_o \times \mathcal{Z}_o \times [0,1]^{d_x} \subseteq \mathbb{R}^{d_{z_o}} \times \mathbb{R}^{d_{z_h}} \times [0,1]^{d_x}$ and is absolutely continuous with respect to Lebesgue's measure.*

Assumption 3.2. *(i) $\inf_{n\geq 1} \inf_{x \in [0,1]^{d_x}} \sigma_{min}(\mathbb{E}\left[\rho(Z, \Pi_n h_0)\rho(Z, \Pi_n h_0)^t \mid X = x\right]) > 0$. (ii) For some $p \geq 4$ and for every $M_0 > 0$*

$$\sup_{x \in [0,1]^{d_x}} \mathbb{E}\left[\sup_{h \in \mathcal{H}^{M_0}} \|\rho(Z,h)\|^p \mid X = x \right] < +\infty \quad \textit{and} \quad \sup_{n \geq 1} \mathbb{E}\left[\max_{1 \leq i \leq n} \sup_{h \in \mathcal{H}^{M_0}} \|\rho(Z_i,h)\|^p \right] < +\infty.$$

Assumption 3.1 rules out time-series framework and places restrictions on the support of P. We do not restrict the support of P^{Z_h} which means that we can accommodate both cases where P^{Z_h} has unbounded support or where P^{Z_h} has bounded support with a Lebesgue density bounded away from zero or not. In the context of GMM estimation, Assumption 3.2 would be replaced by $\sup_{x \in [0,1]^{d_x}} \mathbb{E}\left[\sup_{h \in \mathcal{H}^{M_0}} \|\rho(Z,h)\|^p \mid X = x\right] < +\infty$ for every $M_0 > 0$ and some $p \geq 2$. We need stronger conditions for our GEL procedure in order to perform a linear expansion of $\widehat{\mathcal{L}}_n(h)$ with explicit remainder and control this remainder uniformly over $h \in \mathcal{H}^{M_0}$ for every $M_0 > 0$. As will be apparent in the proofs, the objective we minimize for GEL estimators is actually similar to the objective in (3.2) up to a residual term that proves challenging to control.

Let $\Pi_n h_0(\cdot) = \sum_{k=1}^{\varphi(b_n^{-1})} \langle h_0, q_k \rangle_{leb} q_k(\cdot)$ with $\langle h_0, q_k \rangle_{leb} = \int_{\mathcal{Z}_h} h_0(z_h) q_k(z_h) dz_h$. Remark that for every $n \geq 1$, $\Pi_n h_0 \in \mathcal{H}_n$ and by construction $\|h_0 - \Pi_n h_0\|_{L_2(leb)} = o(1)$. The next assumption recalls the definition of $\mathcal{H}$ and the identification condition (3.1) and places a restriction on $(\Pi_n h_0)_{n\geq 1}$ and P^{Z_h}:

Assumption 3.3. *(i) $\mathcal{H} := \{h \in L_2(P) : \|h\|_{m,2} < +\infty\}$. (ii) For every $h \in \mathcal{H}$, $\mathbb{E}\left[\rho(Z,h) \mid X\right] = 0$ P^X-a.s $\iff \|h - h_0\|_2 = 0$. (iii) $Pen(\cdot) := \|\cdot\|_{m,2}^2$, $Pen(h_0) < +\infty$ and $|Pen(h_0 - \Pi_n h_0)| = O(1)$. (iv) P^{Z_h} has Lebesgue density f_{Z_h} uniformly bounded from above, $m - d_{z_h}/2 > 0$ and either $\mathbb{E}\left[\langle Z_h \rangle^\gamma\right] < +\infty$ for some $\gamma > 0$ or $\mathcal{Z}_h = [0,1]^{d_{z_h}}$.*

[4] If we choose $\mathcal{H} := \{h \in L_2(P) : \|h\|_{2,m} \leq M\}$ for some fixed $M < +\infty$, then it is possible to achieve regularization via the sieve dimension only.

The last condition of Assumption 3.3(iii) requires that $\Pi_n h_0$ be within controlled distance of h_0 in $\|\cdot\|_{m,2}$-norm. Observe that $\|h_0 - \Pi_n h_0\|_{L_2(leb)} = o(1)$ is not enough to ensure $|Pen(h_0 - \Pi_n h_0)| = O(1)$ since $\|h_0 - \Pi_n h_0\|_{L_2(leb)} \leq \|h_0 - \Pi_n h_0\|_{m,2}$. Assumption 3.3(iii) also has a simple but useful consequence: for every $M > 0$, there exists $N_M \geq 1$ such that for every $n > N_M$

$$Pen\left(\Pi_n h_0\right) \leq 2(Pen\left(h_0 - \Pi_n h_0\right) + Pen\left(h_0\right)) \leq M + 2Pen\left(h_0\right) < +\infty.$$

This implies that for every $n \geq 1$

$$Pen\left(\Pi_n h_0\right) \leq \max_{1 \leq m \leq N_M} Pen\left(\Pi_m h_0\right) + M + 2Pen\left(h_0\right) < +\infty.$$

As a result, we can assume without loss of generality that for every $n \geq 1$, $\Pi_n h_0$ belongs to $\mathcal{H}^{M_0}$ for some $M_0 > 0$ independent from n.

We explained in Section 3.2.2 that it is essential to have $\|h\|_2 \leq C\|h\|_{2,m}$ for some $C > 0$ and every $h \in \mathcal{H}$ to ensure that the choice $Pen(h) = \|h\|_{2,m}^2$ is helpful to regularize the estimation problem. As shown in Lemma 3.9, Assumption 3.3(iv) entails that the penalty function $Pen(\cdot)$ is *precompact*, *i.e* for every $M_0 > 0$ the closure of $\mathcal{H}^{M_0}$ for the $\|\cdot\|_2$-norm (that we denote $\overline{\mathcal{H}}^{M_0}$) is a compact subset of $L_2(P)$. This property is key at the end of the consistency proof to bound $\mathbb{E}\left[\|m(X,h)\|^2\right]$ away from zero whenever h is bounded away from h_0 in $\|\cdot\|_2$-norm. Note also that the last condition displayed in Assumption 3.3(iv) is very mild (it does not even impose a moment of order 1 on P^{Z_h}).

Precompact penalties are a generalization of lower semicompact penalties which have been used, *e.g.*, in [37]. Lower semicompact penalties require $\mathcal{H}^{M_0}$ rather than $\overline{\mathcal{H}}^{M_0}$ to be a compact subset of $L_2(P)$ for every $M_0 > 0$. [72, Theorems 1 and 2] give sufficient conditions for $\|\cdot\|_{2,m}$ to be precompact/lower semicompact in $L_2(leb)$. They notably impose $\gamma > d_{z_h}/2$. Interestingly we only need $\gamma > 0$ to obtain that $\|\cdot\|_{2,m}$ be precompact in $L_2(P)$. We emphasize that what really matters is precompactness rather that lower semicompactness.

Let $\widetilde{K}$ and $\widehat{K}$ be two functions that map $\mathbb{R}$ to itself. We define the nonparametric regression kernel as

$$K(x,y,b_n) = \prod_{t=1}^{d_x} \left(\mathbb{1}\left\{0 \leq x^{(t)} < b_n\right\} \widehat{K}\left(\frac{y^{(t)} - x^{(t)}}{b_n}\right) + \mathbb{1}\left\{b_n \leq x^{(t)} \leq 1 - b_n\right\} \widetilde{K}\left(\frac{x^{(t)} - y^{(t)}}{b_n}\right) \right.$$
$$\left. + \mathbb{1}\left\{x^{(t)} > 1 - b_n\right\} \widehat{K}\left(\frac{x^{(t)} - y^{(t)}}{b_n}\right) \right).$$

We need this somewhat complicated kernel to overcome what is called the boundary effect of kernel regression ([75]): when the support of P^X is $[0,1]^{d_x}$, the kernel estimator of $m(x,h)f_X(x)$ is in general not consistent when x lies in what is called the boundary region of $[0,1]^{d_x}$

$$I_{b_n} := \left\{x \in [0,1]^{d_x} : \exists t \in \{1,...,d_x\} \text{ such that } x^{(t)} < b_n \text{ or } x^{(t)} > 1 - b_n\right\}.$$

We present additional restrictions on $K(\cdot)$ and P under which our kernel estimator has good properties:

Assumption 3.4. *(i) P^X has Lebesgue density $f_X \in C^s([0,1]^{d_x})$ with uniformly bounded partial derivatives of order s and $0 < \underline{f}_X \leq f_X(x) \leq \overline{f}_X < +\infty$.*

(ii) For every $l \in \{1,...,q\}$, every $M_0 > 0$ and every $h \in \mathcal{H}^{M_0}$, $m_l(\cdot,h) \in C^s([0,1]^{d_x})$ and for every $\beta : |\beta| = s$, $\sup_{(x,h) \in [0,1]^{d_x} \times \mathcal{H}^{M_0}} \left|D^\beta m_l(x,h)\right| < +\infty$.

(iii) $\widetilde{K}(\cdot)$ is bounded, has support $[-1,1]$ and satisfies $\int_{-1}^1 \widetilde{K}(u)du = 1$ and $\int_{-1}^1 u^j \widetilde{K}(u)du = 0$ for every $j \in \{1,...,s-1\}$.

(iv) $\widehat{K}(\cdot)$ is bounded, has support $[0,1]$ and satisfies $\int_0^1 \widehat{K}(u)du = 1$ and $\int_0^1 u^j \widehat{K}(u)du = 0$ for every $j \in \{1,...,s-1\}$.

(v) The classes of functions $\left\{\widetilde{K}\left(\frac{x-\cdot}{b}\right) : (x,b) \in [0,1]\times\mathbb{R}_+^*\right\}$ *and* $\left\{\widehat{K}\left(\frac{x-\cdot}{b}\right) : (x,b) \in [0,1]\times\mathbb{R}_+^*\right\}$ *are VC-type for constants that do not depend on* n*.*

To show consistency of a kernel estimator, we need to take care of two terms: a variance part and a bias part ([135]). Assumptions 3.4(i) to (iv) are made to control the bias induced by kernel regression. Assumptions 3.4(i)-(ii) allow to make a Taylor-Lagrange expansion of $m(\cdot,h)f_X(\cdot)$ with explicit remainder of order s at every point $x \in [0,1]^{d_x}$ and to control this remainder uniformly in $(x,h) \in [0,1]^{d_x}\times\mathcal{H}^{M_0}$. Similar restrictions can be found in [101]. A more primitive version of Assumption 3.4(ii) would require additional smoothness of the Lebesgue density of the distribution $P^{Z|X}$. Under Assumptions 3.4(iii) and (iv), $\widetilde{K}(\cdot)$ (*resp.* $\widehat{K}(\cdot)$) is a univariate kernel of order s (*resp.* a univariate boundary kernel of order s) so that all bias terms that converge to 0 slower than b_n^s disappear. Assumptions 3.4(iii) and (iv) can be found in [86], [125] and [48].

Assumption 3.4(v) helps to control the "variance" part uniformly in $x \in [0,1]^{d_x}$ when estimating $m(x,\Pi_n h_0)f_X(x)$. This is specific to our GEL estimator and would not be needed in the GMM approach. A class of functions $\mathcal{F}$ with envelope function F is called VC-type for positive constants A and v if for every ϵ, $\sup_Q N(\epsilon\|F\|_{L_2(Q)},\mathcal{F},\|\cdot\|_{L_2(Q)}) \leq (A/\epsilon)^v$ where the supremum is taken over all finitely supported probability mesures (see [43] for more details). A and v are in general allowed to depend on n but we rule this out with our assumption. The advantage of Assumption 3.4(v) is that it does not require $\widetilde{K}(\cdot)$ and $\widehat{K}(\cdot)$ to be continuous. Following discussions in [115], [77] and [78], the following popular kernels exhibit the VC-type property: uniform, triangular, Epanechnikov, biweight, triweight and (truncated) Gaussian. This assumption could be replaced with a smoothness assumption on the kernel (as in [107]).

Assumption 3.5. *(i) For every* $M_0 > 0$ *the map* $h \mapsto \mathbb{E}\left[\|m(X,h)\|^2\right]$ *is lower semicontinuous on* $\mathcal{H}^{M_0}$ *with respect to the* $L_2(P)$ *norm. (ii)* $\gamma > m - d_{z_h}/2 > 0$ *if* $\mathcal{Z}_h$ *is unbounded. (iii) For every* $M_0 > 0$*, there exists a constant* L *such that for every* $(h_1,h_2) \in \mathcal{H}^{M_0}\times\mathcal{H}^{M_0}$ $||m(X,h_1) - m(X,h_2)|| \leq L\,||h_1 - h_2||_{\infty,\gamma}$ $P^X - a.s.$

Assumption 3.5 is useful in combination with Assumption 3.3(iv) to prove that the map $h \mapsto \mathbb{E}\left[\|m(X,h)\|^2\right]$ is bounded away from zero whenever h is bounded away from h_0 in $\|\cdot\|_2$-norm. We show that Assumptions 3.5(i) and (iii) are satisfied in the NPIV and NPQIV cases under simple low-level conditions.

Example 3.1 (NPIV)**.** *In the NPIV we have* $\rho(Z,h) = Z_o - h(Z_h)$*,* $d_z = 2$ *and* $d = 1$*. Hence,*

$$\begin{aligned}
&\left|\,\mathbb{E}\left[\|m(X,h_1)\|^2\right] - \mathbb{E}\left[\|m(X,h_2)\|^2\right]\,\right| \\
\leq&\sqrt{\mathbb{E}\left[(\|m(X,h_1)\| + \|m(X,h_2)\|)^2\right]}\sqrt{\mathbb{E}\left[(\|m(X,h_1)\| - \|m(X,h_2)\|)^2\right]} \\
\leq&\sqrt{\mathbb{E}\left[8Z_o^2 + 4h_1(Z_h)^2 + 4h_2(Z_h)^2\right]}\|h_1 - h_2\|_2 \\
\leq&(\sqrt{8\mathbb{E}\left[Y^2\right]} + K)\|h_1 - h_2\|_2,
\end{aligned}$$

where $K = 2\max\left\{1, \sqrt{\sup_{z_h\in\mathcal{Z}_h} f_{Z_h}(z_h)}\right\}\sqrt{M_0}$*. What is more*

$$\begin{aligned}
\|m(X,h_1) - m(X,h_2)\| \leq&\mathbb{E}\left[\,|h_2(Z_h) - h_1(Z_h)|\,\mid X\right] \\
\leq&\|h_2 - h_1\|_{\infty,\gamma}\sup_{x\in[0,1]^{d_x}}\mathbb{E}\left[\langle Z_h\rangle^\gamma \mid X = x\right].
\end{aligned}$$

If $\sup_{z_h\in\mathcal{Z}_h} f_{Z_h}(z_h) < +\infty$ *and* $\sup_{x\in[0,1]^{d_x}}\mathbb{E}\left[\langle Z_h\rangle^\gamma \mid X = x\right] < +\infty$*, Assumptions 3.5(i) and (iii) are true (we can pick* $L = \sup_{x\in[0,1]^{d_x}}\mathbb{E}\left[\langle Z_h\rangle^\gamma \mid X = x\right]$ *in (iii)).* □

Example 3.2 (NPQIV)**.** *In the NPQIV we have* $\rho(Z,h) = \mathbb{1}\left\{Z_o \leq h(Z_h)\right\} - \tau$ *for some* $\tau \in (0,1)$*. Hence,*

by using the Law of Iterated Expectations

$$\begin{aligned}
&\|m(X,h_1)-m(X,h_2)\| = |\mathbb{E}\left[\mathbb{1}\left\{Z_o \leq h_1(Z_h)\right\} - \mathbb{1}\left\{Z_o \leq h_2(Z_h)\right\} \mid X\right]| \\
=&\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left\{Z_o \leq h_1(Z_h)\right\} - \mathbb{1}\left\{Z_o \leq h_2(Z_h)\right\} \mid X, Z_h\right] \mathbb{1}\left\{h_1(Z_h) \geq h_2(Z_h)\right\} \mid X\right] \\
&+ \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left\{Z_o \leq h_2(Z_h)\right\} - \mathbb{1}\left\{Z_o \leq h_1(Z_h)\right\} \mid X, Z_h\right] \mathbb{1}\left\{h_2(Z_h) > h_1(Z_h)\right\} \mid X\right] \\
\leq& \sup_{(z,x)\in\mathcal{Z}\times[0,1]^{d_x}} f_{Z_o|X,Z_h}(z_o \mid x, z_h)\mathbb{E}\left[\,|h_1(Z_h)-h_2(Z_h)| \mid X\right] \\
\leq& \sup_{(z,x)\in\mathcal{Z}\times[0,1]^{d_x}} f_{Z_o|X,Z_h}(z_o \mid x, z_h) \sup_{x\in[0,1]^{d_x}} \mathbb{E}\left[\langle Z_h\rangle^{\gamma} \mid X=x\right] \|h_1-h_2\|_{\infty,\gamma}.
\end{aligned}$$

If $\sup_{x\in[0,1]^{d_x}} \mathbb{E}\left[\langle Z_h\rangle^{\gamma} \mid X=x\right] < +\infty$ *and* $\sup_{(z,x)\in\mathcal{Z}\times[0,1]^{d_x}} f_{Z_o|X,Z_h}(z_o \mid x, z_h) < +\infty$*, we can pick* $L = \sup_{(z,x)\in\mathcal{Z}\times[0,1]^{d_x}} f_{Z_o|X,Z_h}(z_o \mid x, z_h) \sup_{x\in[0,1]^{d_x}} \mathbb{E}\left[\langle Z_h\rangle^{\gamma} \mid X=x\right]$ *in Assumption 3.5(iii). Following the NPIV example and the previous proof, we can claim that* $\sup_{(z,x)\in\mathcal{Z}\times[0,1]^{d_x}} f_{Z_o|X,Z_h}(z_o \mid x, z_h) < +\infty$ *is enough to verify Assumption 3.5(i).* □

It is convenient to introduce several classes of functions:

1. $\mathcal{F}^{l}_{n,x_1} := \left\{(x,z) \mapsto K\left(\frac{x_1-x}{b_n}\right)\rho_l(z,h) : h \in \mathcal{H}^{M_0}\right\}$.

2. $\mathcal{F}^{l,l'}_{n,x_1} := \left\{(x,z) \mapsto K\left(\frac{x_1-x}{b_n}\right)\rho_l(z,h)\rho_{l'}(z,h) : h \in \mathcal{H}^{M_0}\right\}$.

For every $(x_1,l) \in [0,1]^{d_x} \times \{1,...,d\}$, $\mathcal{F}^{l}_{n,x_1}$ admits an envelope $F^{l}_{n,x_1} : (x,z) \mapsto |K(x_1,x,b_n)| \times (\sup_{h\in\mathcal{H}^{M_0}} \|\rho(z,h)\| + 1)$. Similarly, $\mathcal{F}^{l,l'}_{n,x_1}$ admits an envelope $F^{l,l'}_{n,x_1} : (x,z) \mapsto$ $|K(x_1,x,b_n)| \times \left(\sup_{h\in\mathcal{H}^{M_0}} \|\rho(z,h)\|^2 + 1\right)$. In the following assumption, we let $\mathcal{G}_{n,x_1}$ (*resp.* G_{n,x_1}) stand for $\mathcal{F}^{l}_{n,x_1}$ or $\mathcal{F}^{l,l'}_{n,x_1}$ (*resp.* F^{l}_{n,x_1} or $F^{l,l'}_{n,x_1}$).

Assumption 3.6. *For every* $M_0 > 0$ [5]

$$\sup_{x_1\in[0,1]^{d_x}} \int_0^1 \sqrt{1+\log N_{[\,]}\left(\epsilon \left|\left|G_{n,x_1}(X,Z)\right|\right|_{L_2(P^{X,Z})}, \mathcal{G}_{n,x_1}, L_2\left(P^{X,Z}\right)\right)} d\epsilon < +\infty,$$

Assumption 3.6 is a high-level condition which imposes that for every $l \in \{1,...,d\}$ (*resp.* every $(l,l') \in \{1,...,d\}^2$) the class $\mathcal{F}^{l}_{n,x_1}$ (*resp.* $\mathcal{F}^{l,l'}_{n,x_1}$) is not too complex in terms of entropy.

The NPIV model is nested in what can be called the class of "Lipschitz-in-parameter" models, namely models in which $|\rho_l(Z,h_1) - \rho_l(Z,h_2)| \leq L(Z)\,|h_1(Z_h) - h_2(Z_h)|$ for every $l \in \{1,...,d\}$ and $(h_1,h_2) \in \mathcal{H}\times\mathcal{H}$. In Lemma 3.13, we show that Assumption 3.6 is verified in the NPQIV model and in the class of Lipschitz models under the following condition

Assumption 3.7. *(i) Lipschitz case:* $\sup_{x\in[0,1]^{d_x}} \mathbb{E}\left[L(Z)^4\langle Z_h\rangle^{4\gamma} \mid X=x\right] < +\infty$ *for every* $M_0 > 0$ *and some* γ *such that* $\gamma > m - \frac{d_{z_h}}{2} > 0$*;* $\sup_{x\in[0,1]^{d_x}} \mathbb{E}\left[\sup_{h\in\mathcal{H}^{M_0}} \|\rho(Z,h)\|^4 \mid X=x\right] < +\infty$*. (ii) NPQIV case:* $\sup_{x\in[0,1]^{d_x}} \mathbb{E}\left[\langle Z_h\rangle^{\gamma} \mid X=x\right] < +\infty$ *for some* $\gamma > 0$ *such that* $\gamma > m - \frac{d_{z_h}}{2} > 0$*;* $m/d_{z_2} > 1$*;* $\sup_{(z,x)\in\mathcal{Z}\times[0,1]^{d_x}} f_{Z_o|X,Z_h}(z_o \mid x, z_h) < +\infty$.

We are now in a position to prove the consistency of $\widehat{h}_n$.

Theorem 3.1. *Under Assumptions 3.1 to 3.6 and* $\max\left\{\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}}, b_n^s, \|\Pi_n h_0 - h_0\|_{\infty,\gamma}\right\} = O(\sqrt{\alpha_n}) = o(n^{-1/p})$

$$\|\widehat{h}_n - h_0\|_2 = o_P(1).$$

[5] Note that the bound cannot be uniform in M_0. The bound diverges when $M_0 \to +\infty$.

This theorem is the equivalent of Theorem 3.2 in [37]. The restriction $\max\left\{\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}}, b_n^s, \|\Pi_n h_0 - h_0\|_{\infty,\gamma}\right\} = O(\sqrt{\alpha_n})$ is equivalent to the condition (13) in the statement of Theorem 3.2 in [37]. The terms $\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}}$ and b_n^s stem from controlling the variance and bias terms when estimating the weighted conditional expectation operator $\varphi \mapsto \mathbb{E}\left[\varphi(Z) \mid X = \cdot\right] f_X(\cdot)$ and $\|\Pi_n h_0 - h_0\|_{\infty,\gamma}$ is the approximation error due to the use of $\mathcal{H}_n$ instead of $\mathcal{H}$. With the GMM procedure, $\|\Pi_n h_0 - h_0\|_{\infty,\gamma}$ would be replaced with $\|\Pi_n h_0 - h_0\|_2$. We impose an additional condition, namely $O(\sqrt{\alpha_n}) = o(n^{-1/p})$ which is not needed in [37]. This condition is introduced to get a sharp upper bound on the GEL criterion at $\Pi_n h_0$. When will this condition be satisfied? To start with, we can use the fact that

$$\max\left\{\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}}, b_n^s, \|\Pi_n h_0 - h_0\|_{\infty,\gamma}\right\} \leq \max\left\{\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}}, b_n^{s\wedge m}, \|\Pi_n h_0 - h_0\|_{\infty,\gamma}\right\},$$

and take $\sqrt{\alpha_n}$ of the order of the upper bound. Secondly b_n must be chosen to balance the terms $\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}}$ and $b_n^{s\wedge m}$ to minimize $\max\left\{\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}}, b_n^{s\wedge m}\right\}$ ([135]). Since $|\log b_n| \leq C \log n$ here (as can be seen from the assumption $\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}} = o(n^{-1/p}) = o(1)$), we can pick $b_n = (C \log n/n)^{1/(2(s\wedge m)+d_x)}$. This choice implies $\max\left\{\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}}, b_n^{s\wedge m}\right\} \leq (C \log n/n)^{s\wedge m/(2(s\wedge m)+d_x)}$ which goes to zero faster than $n^{-1/p}$ as long as $s \wedge m > d_x/(p-2)$. In the case where $\mathcal{Z}_h = [0,1]^{d_{z_h}}$, γ can be chosen equal to zero and we get $\|\Pi_n h_0 - h_0\|_{\infty,\gamma} = O(\varphi(b_n)^{-m/d_{z_h}})$ when $(q_k(\cdot))_{k\geq 1}$ is a tensor polynomial basis of $L_2(leb)$ ([37]). Any $\varphi(\cdot) \geq \lceil\cdot\rceil^{d_{z_h}}$ satisfies for some $\bar{c}$, $\varphi(b_n)^{-m/d_{z_h}} \leq \bar{c} b_n^m \leq \bar{c} b_n^{s\wedge m}$.

As mentioned in Section 3.2.2 we could formulate a consistency result for the estimator h_n^* (which is found by optimizing the GEL criterion over the whole of $\mathcal{H}$). In all our assumptions, $\Pi_n h_0$ would have to be replaced with h_0 and the term $\|\Pi_n h_0 - h_0\|_{\infty,\gamma}$ in the statement of Theorem 3.1 would vanish.

3.3.2 Rate

We want to derive (an upper bound on) the consistency rate of $\mathbb{E}\left[\|m(X, \widehat{h}_n)\|^2\right]$ to 0. As detailed after the statement of the theorem, this result is crucial in order to derive a rate of convergence of $\widehat{h}_n$ to h_0 in $\|\cdot\|_2$-norm. This last result however is beyond the scope of this article.

Under exactly the same assumptions as those of Theorem 3.1, we obtain the first result of Theorem 3.2, namely a sub-optimal upper bound on $\mathbb{E}\left[\|m(X, \widehat{h}_n)\|^2\right]$. To improve upon this sub-optimal result, we impose more stringent moment conditions:

Assumption 3.8. *For every $\epsilon > 0$ and $M_0 > 0$, $\mathbb{E}\left[\exp\left(\epsilon \sup_{h\in\mathcal{H}^{M_0}} \|\rho(Z,h)\|\right)\right] < +\infty$.*

This assumption is automatically satisfied if $\rho(\cdot)$ is uniformly bounded from above which is true in the NPQIV model. In the NPIV model, the assumption is valid if Z_o is sub-Gaussian and $\langle Z_h\rangle^\gamma$ is sub-Gaussian are almost-surely bounded. For a definition of sub-Gaussian random variables, we refer the reader to [25]. Broadly speaking, any continuous distribution that has tails of the order of a Gaussian one is sub-Gaussian. Our assumption would also allow for distributions with tails slightly fatter than Gaussian ones.

Theorem 3.2. *Let $\nu_n = \max\left\{\sqrt{\frac{|\log b_n|}{nb_n^{d_x\vee d_{z_h}}}}, b_n^{s\wedge m}, \|\Pi_n h_0 - h_0\|_{\infty,\gamma}\right\}$. Under Assumptions 3.1 to 3.6 and $\nu_n = O(\sqrt{\alpha_n}) = o(n^{-1/p})$*

$$\mathbb{E}\left[\|m(X, \widehat{h}_n)\|^2\right] = O_P\left(\nu_n \vee \sqrt{\alpha_n}\right).$$

If Assumption 3.2(ii) is replaced with Assumption 3.8

$$\mathbb{E}\left[\|m(X, \widehat{h}_n)\|^2\right] = O_P\left((\nu_n^2 \vee \alpha_n)\log n\right).$$

We denote the first part of the theorem as a *slow rate result* and the second one as a *fast rate result*. To obtain the slow rate of convergence, the proof is very close to that of Theorem 3.1. To obtain the fast rate, we use an iterative argument that bears similarities with Lemma 3 in [127]. Picking $\alpha_n \asymp \max\left\{\sqrt{\frac{|\log b_n|}{nb_n^{d_x \vee d_{z_h}}}}, b_n^{s\wedge m}, \|\Pi_n h_0 - h_0\|_{\infty,\gamma}\right\}$ and assuming $\|\Pi_n h_0 - h_0\|_{\infty,\gamma} = O(b_n^m)$ (see the discussion after Theorem 3.1 for an example where this holds), we obtain by balancing terms that

$$\sqrt{\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right]} = O\left(\sqrt{\log n}\left(\frac{\log n}{n}\right)^{\frac{s\wedge m}{2(s\wedge m)+d_x\vee d_{z_h}}}\right) = O\left(n^{-\frac{s\wedge m}{2(s\wedge m)+d_x\vee d_{z_h}}}\log n\right).$$

We also used $(\log n)^{\frac{s\wedge m}{2(s\wedge m)+d_x\vee d_{z_h}}} \leq \sqrt{\log n}$. This rate is up to a log term the minimax rate of estimation in $L_2(leb)$ risk in nonparametric regression when the dimension of the regressors is $d_x \vee d_{z_h}$ and the smoothness of the regression function is $s \wedge m$. This is not surprising since: (i) we estimate a weighted conditional expectation operator $\varphi \mapsto \mathbb{E}\left[\varphi(Z) \mid X = \cdot\right] f_X(\cdot)$ with smoothness s and where the dimension of the conditioning variable is d_x; (ii) h has smoothness m and its domain is a subset of $\mathbb{R}^{d_{z_h}}$. It is not obvious to compare our rate results with those of [40] as those authors derive directly a consistency rate of $\widehat{h}_n$ to h_0 in $\|\cdot\|_{L_2(leb)}$-norm. Their result is stronger than ours but limited to the NPQIV model. Note though that their rate involves the quantity $\max\left\{\frac{\sqrt{J_n}}{n^{1/4}}, \alpha_n^{1/4}\right\}$ where J_n is the equivalent of $b_n^{-d_x\vee d_{z_h}}$. We can see that the condition $\frac{\sqrt{J_n}}{n^{1/4}} = o(1)$ is stronger than $\sqrt{\frac{|\log b_n|}{nb_n^{d_x\vee d_{z_h}}}} = o(1)$.

With the GMM approach, we could obtain directly the fast rate without additional moment assumptions and as explained earlier p in Assumption 3.2 could be chosen equal to 2. We could even get rid of $|\log b_n|$ in the term $\frac{|\log b_n|}{nb_n^{d_x\vee d_{z_h}}}$. It remains an open question whether moment conditions could be weakened with GEL estimators to obtain fast rates of convergence.

The rate result for $\sqrt{\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right]}$ is not directly useful. However it is one of the two components to derive consistency of $\widehat{h}_n$ to h_0 in $\|\cdot\|_2$-norm and then asymptotic normality of plug-in estimates of functionals of h_0. We explain how $\sqrt{\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right]}$ impacts the convergence rate in $\|\cdot\|_2$. To do so, we follow the general exposition in [37]. The first requirement is to find a norm $\|\cdot\|_W$ such that for every M_0, there exists $\bar{c}$ such that $\|h-h_0\|_W^2 \leq \bar{c}\min\left\{\mathbb{E}\left[\|m(X,h)\|^2\right], \|h-h_0\|_2^2\right\}$ uniformly over $\mathcal{H}^{M_0}$. As explained in [37], this norm can be chosen equal to $\sqrt{\mathbb{E}\left[\mathbb{E}\left[h(Z_h) \mid X\right]^2\right]}$ in the NPIV. Then we can write

$$\begin{aligned}\|\widehat{h}_n - h_0\|_2 \leq& \|\widehat{h}_n - \Pi_n h_0\|_2 + \|\Pi_n h_0 - h_0\|_2 \\ \leq& \frac{\|\widehat{h}_n - \Pi_n h_0\|_2}{\|\widehat{h}_n - \Pi_n h_0\|_W} \times \|\widehat{h}_n - \Pi_n h_0\|_W + \|\Pi_n h_0 - h_0\|_2.\end{aligned}$$

It is shown in Section 3.6 that $\widehat{h}_n \in \mathcal{H}^{M_0}$ for some $M_0 > 0$ ***w.p.a.1***. As a result, the following inequality is valid ***w.p.a.1***

$$\begin{aligned}\|\widehat{h}_n - h_0\|_2 \leq& \underbrace{\sup_{h\in\mathcal{H}_n^{M_0}:\|h-\Pi_n h_0\|_W\neq 0} \frac{\|h-\Pi_n h_0\|_2}{\|h-\Pi_n h_0\|_W}}_{=:\tau_n} \times \|\widehat{h}_n - \Pi_n h_0\|_W + \|\Pi_n h_0 - h_0\|_2 \\ \leq& \tau_n\left(\|\widehat{h}_n - h_0\|_W + \|\Pi_n h_0 - h_0\|_W\right) + \|\Pi_n h_0 - h_0\|_2 \\ \leq& \sqrt{\bar{c}}\,\tau_n\left(\sqrt{\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right]} + \|\Pi_n h_0 - h_0\|_2\right) + \|\Pi_n h_0 - h_0\|_2.\end{aligned}$$

The quantity τ_n is called the *sieve measure of local ill-posedness* in [37]. As the authors explain, τ_n goes to infinity in general in models that satisfy (3.1). The speed at which τ_n explodes depends on $\varphi(b_n^{-1})$ so

that $\varphi(b_n^{-1})$ cannot be chosen arbitrarily large. Criteria similar to τ_n but that still allow to select $\varphi(b_n^{-1})$ very large exist ([37]). A vast literature has studied generic conditions to control τ_n or equivalent criteria: the most popular conditions are the source and Hilbert scale conditions (see [39] and [33] for an extensive treatment of this question).

3.4 Conclusion

Allowing for endogenous regressors in the nonparametric mean or quantile regression models is a challenging issue that has initiated a vast literature in theoretical econometrics. The NPIV and NPQIV are two instances of econometric problems that turn out to be ill-posed statistical inverse problems. Unlike most articles we resort to a GEL estimation procedure. We show the consistency of our estimator in $L_2(P)$ norm for a wide class of econometric problems that encompasses the NPIV and NPQIV. Our results could be directly combined with well-known arguments that measure the degree of ill-posedness of the problem to obtain the consistency rate of our estimator in $L_2(P)$ norm ([37], [33]). However, we believe that this approach has a major drawback: to the best of our knowledge, only very few parametric families of distributions have been shown to satisfy the so-called source and Hilbert scale conditions which are the two most popular conditions imposed in the literature to control the degree of ill-posedness. One avenue for future research would be to see how much source and Hilbert scale conditions are impacted when we depart "slightly" from the parametric families for which those conditions are verified.

3.5 Proofs of the main results

In this section, we use $\widehat{D}(X_i,h)$ as a shortcut for $\frac{1}{nb_n^{d_x}}\sum_{j=1}^n K_{ij}\rho(Z_j,h)$.

3.5.1 Proof of Theorem 3.1

Throughout the proof, let $\beta_n := \max\left\{\frac{1}{\sqrt{nb_n^{d_x}}}, b_n^s\right\}$ and

$$\zeta_n := \max\left\{\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}}, b_n^s, \|\Pi_n h_0 - h_0\|_{\infty,\gamma}, \sqrt{\alpha_n}, \sqrt{R_n}\right\}.$$

We start by finding a measurable set of large probability on which it is possible to lower bound $\widehat{\mathcal{L}}_n(h)$. As a matter of fact, $\widehat{\mathcal{L}}_n(h)$ is difficult to control since its definition involves a maximisation step based on the entire sample and that does not have a closed form solution. To build the desired set, it is useful to note that by Lemmas 3.3 and 3.11, for every $\epsilon>0$ there exists M_0 and $N_0\geq 1$ such that

$$\mathbb{P}\left(\left\{Pen(\widehat{h}_n)\leq M_0\right\}\cap\left\{\max_{1\leq i\leq n}\sup_{h\in\mathcal{H}^{M_0}}\|\widehat{D}(X_i,h)\|\leq M_0\right\}\right)\geq 1-\epsilon/16 \tag{3.9}$$

for every $n>N_0$. As explained after Assumption 3.3, there exists $M_1>0$ such that $\sup_{n\geq 1} Pen(\Pi_n h_0)\leq M_1$. Without loss of generality, we pick $M_0\geq M_1$ for every $\epsilon>0$.

Let $\mathcal{A}_{1,n}$ stand for $\left\{Pen(\widehat{h}_n)\leq M_0\right\}\cap\left\{\max_{1\leq i\leq n}\sup_{h\in\mathcal{H}^{M_0}}\|\widehat{D}(X_i,h)\|\leq M_0\right\}$ and

$$\mathcal{A}_{2,n} := \left\{\max_{j\in\{1,\dots,n\}}\sup_{(\lambda,h)\in\Lambda_n\times\mathcal{H}^{M_0}}|\psi(\lambda'\rho(Z_j,h))|<+\infty\right\},$$

where $\Lambda_n := \left\{\lambda\in\mathbf{R}^d : \|\lambda\|\leq\zeta_n M_0\right\}$. By Lemma 3.2 ($\mu_n=\zeta_n$ and $C=M_0$), we can claim that the set $\mathcal{A}_{2,n}$ has probability larger than $1-\epsilon/16$ for every $n>N_1$ for some $N_1\geq N_0$. We deduce from this and

(3.9) that for every $n > N_1$

$$\begin{aligned} & 1 - \epsilon/16 \leq \mathbb{P}\left(\mathcal{A}_{1,n}\right) \leq \mathbb{P}\left(\mathcal{A}_{1,n} \cap \mathcal{A}_{2,n}\right) + \frac{\epsilon}{16} \\ \Longrightarrow & \mathbb{P}\left(\mathcal{A}_{1,n} \cap \mathcal{A}_{2,n}\right) \geq 1 - \frac{\epsilon}{8}. \end{aligned}$$

Let

$$A_1 := \max_{1 \leq i \leq n} \frac{1}{n b_n^{d_x}} \sum_{j=1}^{n} |K_{ij}| \sup_{h \in \mathcal{H}^{M_0}} \|\rho(Z_j, h)\|^2 \max_{1 \leq j \leq n} \sup_{(\tau, v, h) \in [0,1] \times \Lambda_n \times \mathcal{H}^{M_0}} \left| \psi''(\tau v^t \rho(Z_j, h)) + 1 \right|$$

and $\mathcal{A}_{3,n} := \left\{ A_1 \leq \frac{M_2}{\epsilon} \right\}$ for some M_2 that depends on M_0 (and therefore on ϵ).

Using Lemmas 3.6 (with $\mu_n = \zeta_n$ and $C = M_0$) and 3.11, we see that for every $\epsilon > 0$, there exists $N_2 \geq N_1$ such that for every $n > N_2$, $\mathcal{A}_{3,n}^c$ has probability at most $\epsilon/8$. Thus

$$\mathbb{P}\left(\mathcal{A}_{1,n} \cap \mathcal{A}_{2,n} \cap \mathcal{A}_{3,n}\right) \geq 1 - \frac{\epsilon}{4} \tag{3.10}$$

Recall now that

$$\Lambda_n(h) := \left\{ \lambda \in \mathbb{R}^d : \psi\left(\lambda' \rho(Z_j, h)\right) \text{ well-defined } \forall j \in \{1, ..., n\} \right\},$$

and for every $n \geq 1$, every $h \in \mathcal{H}$ and every $(\lambda_i)_{i=1}^n \in \Lambda_n(h)^n$

$$\widehat{\mathcal{L}}_n(h) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \sup_{\lambda \in \Lambda_n(h)} \frac{1}{n b_n^{d_x}} \sum_{j=1}^{n} K_{ij} \psi(\lambda^t \rho(Z_j, h)) \quad \geq \quad \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n b_n^{d_x}} \sum_{j=1}^{n} K_{ij} \psi(\lambda_i^t \rho(Z_j, h)).$$

Let $\widehat{\mathcal{L}}_n^L(h) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n b_n^{d_x}} \sum_{j=1}^{n} K_{ij} \psi(-\zeta_n \widehat{D}(X_i, h)^t \rho(Z_j, h))$. We can see that on $\mathcal{A}_{1,n} \cap \mathcal{A}_{2,n} \cap \mathcal{A}_{3,n}$

$$\inf_{h \in \mathcal{H}_n} \left\{ \widehat{\mathcal{L}}_n(h) + \alpha_n Pen\left(h\right) \right\} + R_n \geq \widehat{\mathcal{L}}_n(\widehat{h}_n) + \alpha_n Pen\left(\widehat{h}_n\right) \geq \inf_{h \in \mathcal{H}^{M_0}} \left\{ \widehat{\mathcal{L}}_n^L(h) + \alpha_n Pen\left(h\right) \right\}. \tag{3.11}$$

Consequently, $\mathcal{A}_{1,n} \cap \mathcal{A}_{2,n} \cap \mathcal{A}_{3,n}$ is a set of probability at least $1 - \epsilon/4$ for every $n > N_2$ on which $\widehat{\mathcal{L}}_n(h)$ is lower bounded by a quantity that will prove easier to handle. On the same set, we also have

$$\inf_{h \in \mathcal{H}_n} \left\{ \widehat{\mathcal{L}}_n(h) + \alpha_n Pen\left(h\right) \right\} + R_n \leq \widehat{\mathcal{L}}_n(\Pi_n h_0) + \alpha_n Pen\left(\Pi_n h_0\right) + R_n. \tag{3.12}$$

We know that around $\xi = 0$ the function $\psi(\xi)$ admits a Mean Value (MV) expansion of the form: there exists a $\tau \in (0, 1)$ such that $\psi(\xi) = \psi(0) - \xi + \frac{\xi^2}{2} \psi''(\tau \xi)$. Since on $\mathcal{A}_{1,n} \cap \mathcal{A}_{2,n} \cap \mathcal{A}_{3,n}$, $\max_{(i,j) \in \{1,...,n\}^2} \sup_{h \in \mathcal{H}^{M_0}} \left| \psi(-\zeta_n \widehat{D}(X_i, h)^t \rho(Z_j, h)) \right| < +\infty$, we can use this MV expansion to write for every $h \in \mathcal{H}^{M_0}$.

$$\begin{aligned} \widehat{\mathcal{L}}_n^L(h) =& \frac{\psi(0)}{n} \sum_{i=1}^{n} \frac{1}{n b_n^{d_x}} \sum_{j=1}^{n} K_{ij} + \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n b_n^{d_x}} \sum_{j=1}^{n} K_{ij} \zeta_n \widehat{D}(X_i, h)^t \rho(Z_j, h) \\ & + \frac{1}{2n} \sum_{i=1}^{n} \frac{1}{n b_n^{d_x}} \sum_{j=1}^{n} K_{ij} \zeta_n^2 \left(\widehat{D}(X_i, h)^t \rho(Z_j, h) \right)^2 \psi''(\tau \zeta_n \widehat{D}(X_i, h)^t \rho(Z_j, h)) \\ \geq & \frac{\psi(0)}{n} \sum_{i=1}^{n} \frac{1}{n b_n^{d_x}} \sum_{j=1}^{n} K_{ij} + \frac{\zeta_n}{n} \sum_{i=1}^{n} \|\widehat{D}(X_i, h)\|^2 - \frac{\zeta_n^2}{2n} \sum_{i=1}^{n} \|\widehat{D}(X_i, h)\|^2 A_1 \\ & - \frac{\zeta_n^2}{2n} \sum_{i=1}^{n} \|\widehat{D}(X_i, h)\|^2 \left\| \frac{1}{n b_n^{d_x}} \sum_{j=1}^{n} K_{ij} \rho(Z_j, h) \rho(Z_j, h)^t \right\| \\ \geq & \frac{\psi(0)}{n} \sum_{i=1}^{n} \frac{1}{n b_n^{d_x}} \sum_{j=1}^{n} K_{ij} + \left(1 - \frac{\zeta_n M_2}{2\epsilon} \right) \frac{\zeta_n}{n} \sum_{i=1}^{n} \|\widehat{D}(X_i, h)\|^2 \\ & - \frac{\zeta_n^2}{2n} \sum_{i=1}^{n} \|\widehat{D}(X_i, h)\|^2 \left\| \frac{1}{n b_n^{d_x}} \sum_{j=1}^{n} K_{ij} \rho(Z_j, h) \rho(Z_j, h)^t \right\|. \end{aligned} \tag{3.13}$$

For every $\epsilon > 0$, there exists $N_3 \geq N_2$ such that for every $n > N_3$, $\left(1 - \frac{\zeta_n M_2}{2\epsilon}\right) \geq \frac{1}{2}$. Let $\widetilde{\mathcal{L}}_n^L(h) := \frac{\zeta_n}{2n}\sum_{i=1}^n \|\widehat{D}(X_i,h)\|^2 - \frac{\zeta_n^2}{2n}\sum_{i=1}^n \|\widehat{D}(X_i,h)\|^2 \left\|\frac{1}{nb_n^{d_x}}\sum_{j=1}^n K_{ij}\rho(Z_j,h)\rho(Z_j,h)^t\right\|$. We can infer from our last remark, (3.10), (3.11), (3.12), (3.13) and the fact that $Pen(h) \geq 0$ for every $h \in \mathcal{H}$, that for every $\epsilon > 0$, every $n > N_3$ and every $\eta > 0$

$$\begin{aligned}\mathbb{P}\left(\|\widehat{h}_n - h_0\|_2 \geq \eta\right) &\leq \mathbb{P}\left(\left\{\|\widehat{h}_n - h_0\|_2 \geq \eta\right\} \cap \mathcal{A}_{1,n} \cap \mathcal{A}_{2,n} \cap \mathcal{A}_{3,n}\right) + \frac{\epsilon}{4} \\ &\leq \mathbb{P}\left(\mathcal{A}_{4,n}\right) + \frac{\epsilon}{4}. \end{aligned} \tag{3.14}$$

where

$$\mathcal{A}_{4,n} := \left\{\frac{\psi(0)}{n}\sum_{i=1}^n \frac{1}{nb_n^{d_x}}\sum_{j=1}^n K_{ij} + \inf_{h\in\mathcal{H}^{M_0}:\|h-h_0\|_2\geq\eta} \widetilde{\mathcal{L}}_n^L(h) \leq \widehat{\mathcal{L}}_n(\Pi_n h_0) + \alpha_n P\left(\Pi_n h_0\right) + R_n\right\}.$$

By Assumption 3.5(i), $h \mapsto \mathbb{E}\left[\|m(X,h)\|^2\right]$ is lower semicontinuous on $\mathcal{H}$ for the $L_2(P)$ norm. The set $\{h \in L_2(P) : \|h - h_0\|_2 \geq \eta\}$ is closed in $L_2(P)$ and the set $\overline{\mathcal{H}}^{M_0}$ is compact in $L_2(P)$ as can be seen from Lemma 3.9. As a result, the set $\left\{h \in \overline{\mathcal{H}}^{M_0} : \|h - h_0\|_2 \geq \eta\right\}$ is itself compact in $L_2(P)$ and the map $h \mapsto \mathbb{E}\left[\|m(X,h)\|^2\right]$ attains its lower bound on this set. The identification condition further ensures that this lower bound is strictly positive for every $\eta > 0$. This implies that for every M_0 and η, there exists $\eta^* > 0$ such that the set $\left\{h \in \overline{\mathcal{H}}^{M_0} : \|h - h_0\|_2 \geq \eta\right\}$ is included in $\left\{h \in \overline{\mathcal{H}}^{M_0} : \mathbb{E}\left[\|m(X,h)\|^2\right] \geq \eta^*\right\}$ and then $\left\{h \in \mathcal{H}^{M_0} : \|h - h_0\|_2 \geq \eta\right\} \subseteq \left\{h \in \mathcal{H}^{M_0} : \mathbb{E}\left[\|m(X,h)\|^2\right] \geq \eta^*\right\}$. We can therefore apply Lemma 3.4 with $\mu_n = \zeta_n$, $\mathcal{L}_n^{\mu_n}(h) = \widetilde{\mathcal{L}}_n^L(h)$, $\delta_{1,n} = \eta^*$ and $\delta_{2,n} = +\infty$. From this, Assumption 3.2 and Lemma 3.5, we can claim that for every $\epsilon > 0$ and $\eta > 0$, there exists $N_4 \geq N_3$ such that for every $n > N_4$ and some positive constants M_0, M_3, M_4 and M_5 independent from η and n

$$\begin{aligned}&\mathbb{P}\left(\mathcal{A}_{4,n}\right) + \frac{\epsilon}{4} \\ \leq&\mathbb{P}\left(\mathcal{A}_{4,n} \cap \mathcal{A}_{5,n} \cap \mathcal{A}_{6,n}\right) + \mathbb{P}\left(\mathcal{A}_{5,n}^c\right) + \mathbb{P}\left(\mathcal{A}_{6,n}^c\right) + \frac{\epsilon}{4} \\ \leq&\mathbb{P}\left(\mathcal{A}_{4,n} \cap \mathcal{A}_{5,n} \cap \mathcal{A}_{6,n}\right) + \frac{\epsilon}{2} + \frac{M_4}{n^{d_{z_h}/(d_{z_h}+2m)}} e^{-M_5 n^{d_{z_h}/(d_{z_h}+2m)}}, \end{aligned} \tag{3.15}$$

where

$$\mathcal{A}_{5,n} := \left\{\zeta_n\left(\frac{\underline{f}_X^2}{32}\inf_{h\in\mathcal{H}^{M_0}:\mathbb{E}[\|m(X,h)\|^2]\geq\eta^*} \mathbb{E}\left[\|m(X,h)\|^2\right] - M_3\beta_n^2\right) \leq \inf_{h\in\mathcal{H}^{M_0}:\mathbb{E}[\|m(X,h)\|^2]\geq\eta^*} \widetilde{\mathcal{L}}_n^L(h)\right\},$$

and

$$\mathcal{A}_{6,n} := \left\{\widehat{\mathcal{L}}_n(\Pi_n h_0) + \alpha_n P(\Pi_n h_0) \leq \frac{\psi(0)}{n}\sum_{i=1}^n \frac{1}{nb_n^{d_x}}\sum_{j=1}^n K_{ij} + M_3\zeta_n^2\right\}.$$

The discussion in the previous paragraph enables us to write that

$$\inf_{h\in\mathcal{H}^{M_0}:\mathbb{E}[\|m(X,h)\|^2]\geq\eta^*} \widetilde{\mathcal{L}}_n^L(h) \leq \inf_{h\in\mathcal{H}^{M_0}:\|h-h_0\|\geq\eta} \widetilde{\mathcal{L}}_n^L(h).$$

This, (3.14), (3.15), the definition of $\mathcal{A}_{4,n}$, $\mathcal{A}_{5,n}$ and $\mathcal{A}_{6,n}$ and the fact that $\beta_n \vee \sqrt{R_n} \leq \zeta_n$ and $\frac{M_4}{n^{d_{z_h}/(d_{z_h}+2m)}} e^{-M_5 n^{d_{z_h}/(d_{z_h}+2m)}} = o(1)$ allow us to claim that for every $\epsilon > 0$ and $\eta > 0$, there exist $M_0 > 0$, $M_6 > 0$ and $N_5 \geq N_4$ such that for every $n > N_5$

$$\begin{aligned}&\mathbb{P}\left(\|\widehat{h}_n - h_0\|_2 \geq \eta\right) \leq \mathbb{P}\left(\mathcal{A}_{4,n} \cap \mathcal{A}_{5,n} \cap \mathcal{A}_{6,n}\right) + \epsilon \\ \leq&\mathbb{1}\left\{\inf_{h\in\mathcal{H}^{M_0}:\mathbb{E}[\|m(X,h)\|^2]\geq\eta^*} \mathbb{E}\left[\|m(X,h)\|^2\right] \leq \frac{M_6}{\underline{f}_X^2}\zeta_n\right\} + \epsilon \\ \leq&\mathbb{1}\left\{\eta^* \leq \frac{M_6}{\underline{f}_X^2}\zeta_n\right\} + \epsilon. \end{aligned}$$

Since η^* is strictly positive and does not depend on n, there exists $N \geq N_5$ such that for every $n > N$

$$\mathbb{1}\left\{\eta^* \leq \frac{M_6}{\underline{f}_X^2}\zeta_n\right\} = 0.$$

We conclude that for every $\epsilon > 0$ and $\eta > 0$, there exists $N \geq 1$ such that for every $n > N$

$$\mathbb{P}\left(\|\widehat{h}_n - h_0\|_2 \geq \eta\right) \leq \epsilon.$$

3.5.2 Proof of Theorem 3.2

In a first step, we derive a slow rate of convergence of $\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right]$ to zero. This first step closely follows the proof of Theorem 3.1. In a second step, we improve on the slow rate by adapting an iterative argument presented in Lemma 3 in [127]: in their Lemma 3, the authors show that for a wide class of estimation problems that amount to minimizing an empirical criterion, it is possible to improve the convergence rate of the estimator iteratively under some conditions on the statistical problem at hands.

First step: slow rate of convergence

Let $\zeta_n := \max\left\{\sqrt{\frac{|\log b_n|}{nb_n^{d_x \vee d_{z_h}}}}, b_n^{s\wedge m}, \|\Pi_n h_0 - h_0\|_{\infty,\gamma}, \sqrt{\alpha_n}, \sqrt{R_n}\right\}$ and $\beta_n := \max\left\{\frac{1}{\sqrt{nb_n^{d_x}}}, b_n^s\right\}$. By assumption, $\zeta_n = o(n^{-1/p})$. As a result, the start of the proof of Theorem 3.1 is valid with the new definition of ζ_n. This ensures that for every $\epsilon > 0$, there exist $N_0 \geq 1$ and $M_0 > 0$ such that for every $n > N_0$ and every $r_{1,n} > 0$

$$\begin{aligned}&\mathbb{P}\left(\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right] \geq r_{1,n}^2\right) \leq \mathbb{P}\left(\left\{\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right] \geq r_{1,n}^2\right\} \cap \mathcal{A}_{1,n} \cap \mathcal{A}_{2,n} \cap \mathcal{A}_{3,n}\right) + \frac{\epsilon}{4}\\ \leq&\mathbb{P}\left(\widetilde{\mathcal{A}}_{4,n}\right) + \frac{\epsilon}{4},\end{aligned} \tag{3.16}$$

where $\mathcal{A}_{1,n}$, $\mathcal{A}_{2,n}$ and $\mathcal{A}_{3,n}$ are the same as in the proof of Theorem 3.1 and

$$\begin{aligned}&\widetilde{\mathcal{A}}_{4,n}\\ :=&\left\{\frac{\psi(0)}{n}\sum_{i=1}^n \frac{1}{nb_n^{d_x}}\sum_{j=1}^n K_{ij} + \inf_{h\in\mathcal{H}^{M_0}:\mathbb{E}[\|m(X,h)\|^2]\geq r_{1,n}^2} \widetilde{\mathcal{L}}_n^L(h) \leq \widehat{\mathcal{L}}_n(\Pi_n h_0) + \alpha_n Pen\left(\Pi_n h_0\right) + R_n\right\}.\end{aligned}$$

We assume for now that $r_{1,n}^2 \geq M_1 n^{-2m/(2m+d_{z_h})}$ where $M_1 \geq 1$ corresponds to the constant C_1 in Lemma 3.4. We can therefore apply Lemma 3.4 with $\mu_n = \zeta_n$, $\mathcal{L}_n^{\mu_n}(h) = \widetilde{\mathcal{L}}_n^L(h)$, $\delta_{1,n} = r_{1,n}^2$ and $\delta_{2,n} = +\infty$. From this, Assumption 3.2, Lemma 3.5 and the fact that $\beta_n \leq \zeta_n$, we can claim that for every $\epsilon > 0$, there exist positive constants $(M_i)_{i=0}^4$ such that for every $n > N_0$

$$\begin{aligned}&\mathbb{P}\left(\widetilde{\mathcal{A}}_{4,n}\right) + \frac{\epsilon}{4}\\ \leq&\mathbb{P}\left(\mathcal{A}_{4,n} \cap \widetilde{\mathcal{A}}_{5,n} \cap \mathcal{A}_{6,n}\right) + \frac{\epsilon}{2} + \frac{M_3}{n^{d_{z_h}/(d_{z_h}+2m)}} e^{-M_4 n^{d_{z_h}/(d_{z_h}+2m)}},\end{aligned} \tag{3.17}$$

where

$$\begin{aligned}&\widetilde{\mathcal{A}}_{5,n}\\ :=&\left\{\zeta_n\left(\frac{\underline{f}_X^2}{32}\inf_{h\in\mathcal{H}^{M_0}:\mathbb{E}[\|m(X,h)\|^2]\geq r_{1,n}^2} \mathbb{E}\left[\|m(X,h)\|^2\right] - M_2\zeta_n^2\right) \leq \inf_{h\in\mathcal{H}^{M_0}:\mathbb{E}[\|m(X,h)\|^2]\geq r_{1,n}^2} \widetilde{\mathcal{L}}_n^L(h)\right\},\end{aligned}$$

and $\mathcal{A}_{6,n}$ is the same as in the proof of Theorem 3.1 with M_3 relabelled M_2. As stated in Lemma 3.4, M_3 and M_4 depend on M_0 only. Since the latter depends on ϵ itself, we deduce that M_3 and M_4 ultimately depend on ϵ as well.

We observe that

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{A}_{4,n}\cap\widetilde{\mathcal{A}}_{5,n}\cap\mathcal{A}_{6,n}\right)\\
\leq&\mathbb{1}\left\{\inf_{h\in\mathcal{H}^{M_0}:\mathbb{E}[\|m(X,h)\|^2]\geq r_{1,n}^2}\mathbb{E}\left[\|m(X,h)\|^2\right]\leq\frac{M_5}{\underline{f}_X^2}\zeta_n\right\}\\
\leq&\mathbb{1}\left\{r_{1,n}^2\leq\left(\frac{M_5}{\underline{f}_X^2}\vee 1\right)\zeta_n\right\},
\end{aligned}\tag{3.18}$$

where we used $R_n\leq\zeta_n^2$ and we let $M_5=32(3M_2+2)$.

To conclude, we want to pick $r_{1,n}$ as small as possible such that $r_{1,n}^2\geq M_1 n^{-2m/(2m+d_{z_h})}$ and $\mathbb{1}\left\{r_{1,n}^2\leq\left(\frac{M_5}{\underline{f}_X^2}\vee 1\right)\zeta_n\right\}=0$. Let $M_6:=\left(\frac{M_5}{\underline{f}_X^2}\vee 1\right)$. If we pick $r_{1,n}=2\sqrt{M_1M_6\zeta_n}$, we have $\mathbb{1}\left\{r_{1,n}^2\leq M_6\zeta_n\right\}=0$. We now check that $4M_1M_6\zeta_n\geq M_1n^{-2m/(2m+d_{z_h})}$. Notice that $\max\left\{\frac{1}{nb_n^{d_x\vee d_{z_h}}},b_n^{2(s\wedge m)}\right\}=O(\zeta_n^2)=o(\zeta_n)$ because $b_n=o(1)$ and $\zeta_n=o(1)$. What is more

$$\max\left\{\frac{1}{nb_n^{d_x\vee d_{z_h}}},b_n^{2(s\wedge m)}\right\}\geq\max\left\{\frac{1}{nb_n^{d_{z_h}}},b_n^{2m}\right\}\geq n^{-2m/(2m+d_{z_h})},$$

where the second inequality can be recovered by choosing b_n to balance $\frac{1}{nb_n^{d_{z_h}}}$ and b_n^{2m}. This implies that for n large enough, $4M_1M_6\zeta_n>4M_1\zeta_n^2\geq 4M_1n^{-2m/(2m+d_{z_h})}>M_1n^{-2m/(2m+d_{z_h})}$. We choose $r_{1,n}=2\sqrt{M_1M_6\zeta_n}$ and combine (3.16), (3.17) and (3.18) to claim that for every $\epsilon>0$ there exists $N\geq 1$ and $M>0$ such that for every $n>N$

$$\mathbb{P}\left(\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right]\geq M\zeta_n\right)\leq\epsilon.\tag{3.19}$$

Second step: improved convergence rate

Let $\nu_n=\max\left\{\sqrt{\frac{|\log b_n|}{nb_n^{d_x\vee d_{z_h}}}},b_n^{s\wedge m},\|\Pi_nh_0-h_0\|_{\infty,\gamma},\sqrt{\alpha_n},\sqrt{R_n}\right\}$, $K_n=\lceil\sqrt{\log n}\rceil$, $r_{2,n}=\nu_n^{\frac{1}{2}\sum_{l=0}^{K_n}2^{-l}}\sqrt{\log n}$ and C be a constant greater than 1 to be chosen later. Under Assumption 3.8, the start of the proof of Theorem 3.1 with $\zeta_n=\log n$ remains valid. Consequently, we can claim that for every $\epsilon>0$ there exist $M_0>0$ and $N_0\geq 1$ such that for every $n>N_0$ and every $C>0$

$$\mathbb{P}\left(\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right]\geq C^2r_{2,n}^2\right)\leq\mathbb{P}\left(\left\{\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right]\geq C^2r_{2,n}^2\right\}\cap\mathcal{A}_{1,n}\cap\mathcal{A}_{2,n}\cap\mathcal{A}_{3,n}\right)+\frac{\epsilon}{4}.$$

What is more, (3.19) ensures $\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right]=O_P(\nu_n)=o_P(\nu_n\log n)$. This and Lemma 3.5 imply that for every $\epsilon>0$ and $C\geq 1$, there exists $N_1\geq N_0$ and positive constants M_0 and M_1 independent from C such that for every $n>N_1$

$$\begin{aligned}
&\mathbb{P}\left(\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right]\geq C^2r_{2,n}^2\right)\\
\leq&\mathbb{P}\left(\left\{C^2\nu_n\log n>\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right]\geq C^2r_{2,n}^2\right\}\cap\mathcal{A}_{1,n}\cap\mathcal{A}_{2,n}\cap\mathcal{A}_{3,n}\cap\mathcal{A}_{6,n}\right)+\frac{\epsilon}{2}.
\end{aligned}\tag{3.20}$$

Let $\mathcal{H}_k^{M_0}:=\left\{h\in\mathcal{H}^{M_0}:C^2\nu_n^{\sum_{l=0}^{k-1}2^{-l}}\log n>\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right]\geq C^2\nu_n^{\sum_{l=0}^{k}2^{-l}}\log n\right\}$. Since

$$\begin{aligned}
&\left\{C^2\nu_n\log n>\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right]\geq C^2r_{2,n}^2\right\}\\
=&\bigcup_{k=1}^{K_n}\left\{C^2\nu_n^{\sum_{l=0}^{k-1}2^{-l}}\log n>\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right]\geq C^2\nu_n^{\sum_{l=0}^{k}2^{-l}}\log n\right\},
\end{aligned}$$

we obtain (using also the definition of $(\mathcal{A}_{i,n})_{i=1}^4$)

$$\begin{aligned}
&\mathbb{P}\left(\left\{C^2\nu_n\log n > \mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right] \geq C^2 r_{2,n}^2\right\} \cap \mathcal{A}_{1,n}\cap\mathcal{A}_{2,n}\cap\mathcal{A}_{3,n}\cap\mathcal{A}_{6,n}\right)\\
=&\sum_{k=1}^{K_n}\mathbb{P}\left(\left\{C^2\nu_n^{\sum_{l=0}^{k-1}2^{-l}}\log n > \mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right] \geq C^2\nu_n^{\sum_{l=0}^{k}2^{-l}}\log n\right\} \cap \mathcal{A}_{1,n}\cap\mathcal{A}_{2,n}\cap\mathcal{A}_{3,n}\cap\mathcal{A}_{6,n}\right)\\
\leq&\sum_{k=1}^{K_n}\mathbb{P}\left(\inf_{h\in\mathcal{H}_k^{M_0}}\widetilde{\mathcal{L}}^L(h) \leq M_1\nu_n^2 + R_n\right). \qquad (3.21)
\end{aligned}$$

Let $\beta_n := \max\left\{\frac{1}{\sqrt{nb_n^{d_x}}}, b_n^s\right\}$ and $M_2 \geq 1$ be the constant labelled C_1 in Lemma 3.4. There exists $N_2 \geq N_1$ such that for every $n > N_2$ $\log n > M_2$. As a result, we can claim that for every $n > N_2$ and every $k \geq 1$, $C^2\nu_n^{\sum_{l=0}^{k}2^{-l}}\log n > M_2\nu_n^{\sum_{l=0}^{+\infty}2^{-l}} = M_2\nu_n^2$. The last paragraph in the first step of the proof is enough to check that $\nu_n^2 \geq n^{-2m/(2m+d_{z_h})}$. As a result, we can apply Lemma 3.4 for every $k \in \{1,...,K_n\}$ with $\mu_n = (\log n)^{-1}$, $\delta_{1,n} = C^2\nu_n^{\sum_{l=0}^{k}2^{-l}}\log n$ and $\delta_{2,n} = C^2\nu_n^{\sum_{l=0}^{k-1}2^{-l}}\log n$ to claim that for every M_0 (therefore for every $\epsilon > 0$) there exist positive constants M_3, M_4 and M_5 (independent from k) such that for every $n > N_2$ and every $\eta > 0$

$$\begin{aligned}
&\mathbb{P}\left(\inf_{h\in\mathcal{H}_k^{M_0}}\widetilde{\mathcal{L}}_n(h) \geq \frac{1}{\log n}\left(\frac{\underline{f}_X^2}{32}\inf_{h\in\mathcal{H}_k^{M_0}}\mathbb{E}\left[\|m(X,h)\|^2\right] - A_1\right)\right)\\
\geq&1 - \frac{\eta}{2} - \frac{M_4}{\eta^{d_{z_h}/(d_{z_h}+2m)}}e^{-M_5\eta^{d_{z_h}/(d_{z_h}+2m)}}. \qquad (3.22)
\end{aligned}$$

where $A_1 := M_3\beta_n^2\max\left\{\eta^{-3/2},\eta^{-1}\right\} + \frac{1}{\log n}\frac{M_3}{\sqrt{\eta}}\beta_n\sup_{h\in\mathcal{H}_k^{M_0}}\sqrt{\mathbb{E}\left[\|m(X,h)\|^2\right]}$. By construction $\beta_n \leq \nu_n$ and $C^2\nu_n^{\sum_{l=0}^{k}2^{-l}}\log n \leq \mathbb{E}\left[\|m(X,h)\|^2\right] \leq C^2\nu_n^{\sum_{l=0}^{k-1}2^{-l}}\log n$ for every $h \in \mathcal{H}_k^{M_0}$. We combine this last remark with (3.20), (3.21) and (3.22) and pick $\eta = \epsilon/(4K_n)$ to claim that for every $\epsilon > 0$ and $C \geq 1$, there exists $N_3 \geq N_2$ and positive constants M_0, M_1, M_4, M_5 and M_6 independent from C such that for every $n > N_3$

$$\begin{aligned}
&\mathbb{P}\left(\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right] \geq C^2 r_{2,n}^2\right)\\
\leq&\sum_{k=1}^{K_n}\mathbb{P}\left(\frac{\psi(0)}{n}\sum_{i=1}^{n}\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K_{ij} + \frac{1}{\log n}\left(\frac{\underline{f}_X^2}{32}\inf_{h\in\mathcal{H}_k^{M_0}}\mathbb{E}\left[\|m(X,h)\|^2\right] - A_1\right)\right.\\
&\qquad\left.\leq \inf_{h\in\mathcal{H}_k^{M_0}}\widetilde{\mathcal{L}}^L(h) \leq \frac{\psi(0)}{n}\sum_{i=1}^{n}\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K_{ij} + M_1\nu_n^2 + R_n\right) + \frac{\epsilon}{4}\\
&+ K_n\frac{M_4}{n^{d_{z_h}/(d_{z_h}+2m)}}e^{-M_5 n^{d_{z_h}/(d_{z_h}+2m)}} + \frac{\epsilon}{2}\\
\leq&\sum_{k=1}^{K_n}\mathbb{1}\left\{\frac{1}{\log n}\left(\frac{\underline{f}_X^2}{32}C^2\nu_n^{\sum_{l=0}^{k}2^{-l}}\log n - A_2\right) \leq M_1\nu_n^2 + R_n\right\} + \frac{\epsilon}{4}\\
&+ K_n\frac{M_4}{n^{d_{z_h}/(d_{z_h}+2m)}}e^{-M_5 n^{d_{z_h}/(d_{z_h}+2m)}} + \frac{\epsilon}{2}\\
\leq&\sum_{k=1}^{K_n}\mathbb{1}\left\{\frac{1}{\log n}\left(\frac{\underline{f}_X^2}{32}C^2\nu_n^{\sum_{l=0}^{k}2^{-l}}\log n - A_2\right) \leq (M_1+1)\nu_n^2\right\} + \frac{\epsilon}{4}\\
&+ K_n\frac{M_4}{n^{d_{z_h}/(d_{z_h}+2m)}}e^{-M_5 n^{d_{z_h}/(d_{z_h}+2m)}} + \frac{\epsilon}{2}, \qquad (3.23)
\end{aligned}$$

where $A_2 := M_6\nu_n^2K_n^{3/2} + \frac{M_6\sqrt{K_n}}{\sqrt{\log n}}C\nu_n\nu_n^{\frac{1}{2}\sum_{l=0}^{k-1}2^{-l}} = M_6\nu_n^2K_n^{3/2} + \frac{M_6\sqrt{K_n}}{\sqrt{\log n}}C\nu_n^{\sum_{l=0}^{k}2^{-l}}$.

Let $M_7 := \frac{32}{\underline{f}_X^2}\max\{M_1+1, M_6\}$. For every $k \in \{1,...,K_n\}$

$$\begin{aligned}
&\mathbb{1}\left\{\frac{1}{\log n}\left(\frac{\underline{f}_X^2}{32}C^2\nu_n^{\sum_{l=0}^k 2^{-l}}\log n - A_2\right) \leq (M_1+1)\nu_n^2\right\} \\
\leq &\mathbb{1}\left\{C^2\nu_n^{\sum_{l=0}^k 2^{-l}} \leq M_7\left(\nu_n^2\max\left\{1, \frac{K_n^{3/2}}{\log n}\right\} + \frac{\sqrt{K_n}}{(\log n)^{3/2}}C\nu_n^{\sum_{l=0}^k 2^{-l}}\right)\right\} \\
\leq &\mathbb{1}\left\{\frac{C^2}{2}\nu_n^{\sum_{l=0}^k 2^{-l}} \leq \max\{M_7,1\}\,\nu_n^2\max\left\{1, \frac{K_n^{3/2}}{\log n}\right\}\right\} \\
+ &\mathbb{1}\left\{\frac{C^2}{2}\nu_n^{\sum_{l=0}^k 2^{-l}} \leq \max\{M_7,1\}\,\frac{K_n^{3/2}}{\log n}C\nu_n^{\sum_{l=0}^k 2^{-l}}\right\},
\end{aligned} \tag{3.24}$$

where we used the fact that for positive a, x and y

$$\{x+y \geq a\} \implies \left\{\max\{x,y\} \geq \frac{a}{2}\right\} \iff \left\{x \geq \frac{a}{2}\right\} \cup \left\{y \geq \frac{a}{2}\right\}.$$

We imposed at the beginning of the proof that $K_n = o\left((\log n)^{2/3}\right)$. As a result, for every $\epsilon > 0$, there exists $N_4 \geq N_3$ such that for every $n > N_4$, $\frac{K_n^{3/2}}{\log n} \leq \frac{1}{2\max\{M_7,1\}} \leq 1$ and for every $k \in \{1,...,K_n\}$

$$\begin{aligned}
&\mathbb{1}\left\{\frac{C^2}{2}\nu_n^{\sum_{l=0}^k 2^{-l}} \leq \max\{M_7,1\}\,\nu_n^2\max\left\{1, \frac{K_n^{3/2}}{\log n}\right\}\right\} \\
&+ \mathbb{1}\left\{\frac{C^2}{2}\nu_n^{\sum_{l=0}^k 2^{-l}} \leq \max\{M_7,1\}\,\frac{K_n^{3/2}}{\log n}C\nu_n^{\sum_{l=0}^k 2^{-l}}\right\} \\
\leq &\mathbb{1}\left\{\frac{C^2}{2}\nu_n^{\sum_{l=0}^k 2^{-l}} \leq \max\{M_7,1\}\,\nu_n^2\right\} + \mathbb{1}\left\{\frac{C^2}{2}\nu_n^{\sum_{l=0}^k 2^{-l}} \leq \frac{1}{2}C\nu_n^{\sum_{l=0}^k 2^{-l}}\right\}.
\end{aligned} \tag{3.25}$$

We pick $C = \sqrt{2\max\{M_7,1\}} > 1$. As $K_n = o((\log n)^{2/3})$, $K_n\frac{M_4}{n^{d_{z_h}/(d_{z_h}+2m)}}e^{-M_5 n^{d_{z_h}/(d_{z_h}+2m)}} \leq \frac{\epsilon}{4}$ for n large enough. Combining (3.23), (3.24) and (3.25), we can finally write that for every $\epsilon > 0$, there exists $N \geq 1$ and $M > 0$ such that for every $n > N$

$$\mathbb{P}\left(\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right] \geq Mr_{2,n}^2\right) \leq \epsilon.$$

This is equivalent to $\mathbb{E}\left[\|m(X,\widehat{h}_n)\|^2\right] = O_P(r_{2,n}^2)$. To conclude, we note that for every n

$$r_{2,n} = \nu_n^2 e^{|\log\nu_n|/2^{K_n}}\log n.$$

We know that $\nu_n^2 \geq n^{-2m/(2m+d_{z_h})}$ so that $|\log\nu_n| = O(\log n)$. By definition of K_n, we conclude that $r_{2,n} \sim \nu_n^2\log n$.

3.6 Appendix

In this appendix we use the following additional notations. We let $\widehat{D}(X_i,h) = \frac{1}{nb_n^{d_x}}\sum_{j=1}^n K_{ij}\rho(Z_j,h)$ and $\overline{K} := \max\left\{\sup_{u\in[-1,1]^{d_x}}\left|\widetilde{K}(u)\right|, \sup_{u\in[0,1]^{d_x}}\left|\widehat{K}(u)\right|\right\}$.

We also let for every $\mu_n = o(1)$

$$\mathcal{L}_n^{\mu_n}(h) := \frac{\mu_n}{2n}\sum_{i=1}^n\|\widehat{D}(X_i,h)\|^2 - \frac{\mu_n^2}{2n}\sum_{i=1}^n\|\widehat{D}(X_i,h)\|^2\left\|\frac{1}{nb_n^{d_x}}\sum_{j=1}^n K_{ij}\rho(Z_j,h)\rho(Z_j,h)^t\right\|,$$

and $\mathcal{H}^{M_0}_{\delta_1,\delta_2} = \left\{h \in \mathcal{H}^{M_0} : \delta_1 \leq \mathbb{E}\left[\|m(X,h)\|^2\right] < \delta_2\right\}$ for $0 \leq \delta_1 < \delta_2 \leq +\infty$. Finally, recalling the definitions of $\mathcal{F}^l_{n,x_1}$ and $F^l_{n,x_1}(x,z)$ (*resp.* $\mathcal{F}^{l,l'}_{n,x_1}$ and $F^{l,l'}_{n,x_1}(x,z)$) before Assumption 3.6, we let

$$J_l(n,x_1,M_0,P) := \int_0^1 \sqrt{1+\log N_{[\,]}\left(\epsilon\left|\left|F^l_{n,x_1}(X,Z)\right|\right|_{L_2(P^{X,Z})}, \mathcal{F}^l_{n,x_1}, L_2\left(P^{X,Z}\right)\right)}d\epsilon$$

and

$$J_{l,l'}(n,x_1,M_0,P) := \int_0^1 \sqrt{1+\log N_{[\,]}\left(\epsilon\left|\left|F^{l,l'}_{n,x_1}(X,Z)\right|\right|_{L_2(P^{X,Z})}, \mathcal{F}^{l,l'}_{n,x_1}, L_2\left(P^{X,Z}\right)\right)}d\epsilon.$$

3.6.1 Lemmas

Assumption 3.9. *Let $\varphi(u) = u^p$ for some positive p or $\varphi(u) = e^u$. For every $\epsilon > 0$ and every $M_0 > 0$, $\mathbb{E}\left[\varphi\left(\frac{\sup_{h\in\mathcal{H}^{M_0}}\|\rho(Z,h)\|}{\epsilon}\right)\right] < +\infty$.*

Lemma 3.1. *Let Assumptions 3.1 (i) and 3.9 hold. Then for every $M_0 > 0$*

$$\max_{1\leq i\leq n}\sup_{h\in\mathcal{H}^{M_0}}\|\rho(Z_i,h)\| = o_{a.s}\left(\varphi^{-1}(n)\right).$$

The proof of this lemma follows directly from [101, Lemma D2] and so it is omitted.

Lemma 3.2. *Assume Assumptions 3.1 (i) and 3.9 hold. Let $\Lambda_n := \left\{\lambda \in \mathbb{R}^d : \|\lambda\| \leq C\mu_n\right\}$ where C is a positive constant and μ_n is a positive sequence such that $\mu_n\varphi^{-1}(n) = O(1)$. Then, for every $\psi(\cdot)$ in the GEL family we restrict to and every positive M_0, C and ϵ, there exists $N \geq 1$ such that for every $n > N$*

$$\mathbb{P}\left(\max_{1\leq j\leq n}\sup_{(\lambda,h)\in\Lambda_n\times\mathcal{H}^{M_0}}\left|\psi\left(\lambda^t\rho(Z_j,h)\right)\right| < +\infty\right) \geq 1-\epsilon.$$

Lemma 3.3. *Let $\widehat{h}_n$ be the Pen-EL estimator defined in* (3.8)*. Suppose that $nb_n^{d_x} \to +\infty$, $\max\{\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}}, b_n^s, \|\Pi_n h_0 - h_0\|_{\infty,\gamma}\} = o(n^{-1/p})$ and Assumptions 3.1, 3.2, 3.3, 3.4 and 3.5 (ii) hold. If $\max\left\{\frac{|\log b_n|}{nb_n^{d_x}}, b_n^{2s}, \|\Pi_n h_0 - h_0\|^2_{\infty,\gamma}, R_n\right\} = O(\alpha_n)$, then $Pen(\widehat{h}_n) = O_p(1)$.*

Lemma 3.4. *Let $\beta_n = \max\left\{\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}}, b_n^s\right\}$. Suppose that $\beta_n = o(1)$ and Assumptions 3.1, 3.2(ii), 3.3 and 3.4 hold. Then, we get that for every $M_0 > 0$, there exist $N \geq 1$ and positive constants $C_1 > 1$, C_2, C_3 and C_4 such that for every $n > N$ and every $\epsilon > 0$*

$$\begin{aligned}&\mathbb{P}\left(\inf_{h\in\mathcal{H}^{M_0}_{\delta_{1,n},\delta_{2,n}}}\mathcal{L}_n^{\mu_n}(h) \geq \mu_n\left(\frac{\underline{f}_X^2}{16}\inf_{h\in\mathcal{H}^{M_0}_{\delta_{1,n},\delta_{2,n}}}\mathbb{E}\left[\|m(X,h)\|^2\right] - A_1\right)\right)\\ \geq& 1-\frac{\epsilon}{2}-\frac{C_2}{n^{d_{z_h}/(d_{z_h}+2m)}}e^{-C_3 n^{d_{z_h}/(d_{z_h}+2m)}}.\end{aligned}$$

where $A_1 := C_1\beta_n^2\max\left\{\epsilon^{-3/2},\epsilon^{-1}\right\} + \mu_n\frac{C_1}{\sqrt{\epsilon}}\beta_n\sup_{h\in\mathcal{H}^{M_0}_{\delta_{1,n},\delta_{2,n}}}\sqrt{\mathbb{E}\left[\|m(X,h)\|^2\right]}$ and $\delta_{1,n}$ and $\delta_{2,n}$ are two positive sequences that satisfy $C_1 n^{-2m/(2m+d_{z_h})} \leq \delta_{1,n} < \delta_{2,n} \leq +\infty$.

Lemma 3.5. *Suppose that $nb_n^{d_x} \to +\infty$, $\max\{\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}}, b_n^s, \|\Pi_n h_0 - h_0\|_{\infty,\gamma}\} = o(n^{-1/p})$ for some $p \geq 4$, Assumptions 3.1, 3.2, 3.3, 3.4 and 3.5 (iii) hold. Then uniformly in $i \in \{1,...,n\}$*

$$\begin{aligned}&\sup_{\lambda\in\widehat{\Lambda}(\Pi_n h_0)}\frac{1}{nb_n^{d_x}}\sum_{j=1}^n K_{ij}\psi\left(\lambda^t\rho(Z_j,\Pi_n h_0)\right)\\ \leq&\psi(0)\frac{1}{nb_n^{d_x}}\sum_{j=1}^n K_{ij} + O_P\left(\frac{|\log b_n|}{nb_n^{d_x}} + b_n^{2s} + \|\Pi_n h_0 - h_0\|^2_{\infty,\gamma}\right),\end{aligned}$$

whenever $\psi(\cdot)$ belongs to the GEL family we consider.

Lemma 3.6. *Suppose that Assumptions 3.1 and 3.9 hold. Let* $\Lambda_n := \left\{\lambda \in \mathbb{R}^d : \|\lambda\| \leq C\mu_n\right\}$ *where* C *is a positive constant and* μ_n *is a positive sequence such that* $\mu_n \varphi^{-1}(n) = O(1)$*. Then for every* $\psi(\cdot)$ *in the GEL family we consider and every positive* M_0*,* C *and* ϵ*, there exists* N *such that for every* $n > N$

$$1 - \epsilon \leq \mathbb{P}\left(\max_{1\leq i\leq n} \sup_{(\tau,v,h)\in[0,1]\times\Lambda_n\times\mathcal{H}^{M_0}} \left|\psi^{''}(\tau v^t \rho(Z_i,h)) + 1\right| \leq 1\right).$$

Lemma 3.7. *Suppose that* $nb_n^{d_x} \to +\infty$ *and Assumptions 3.1, 3.2(ii), 3.3, 3.4 and 3.6 hold. Then there exists* $N \geq 1$ *such that for every* $n > N$ *and every* $M_0 > 0$

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \sup_{h\in\mathcal{H}^{M_0}} \left|\left|\widehat{D}(X_i,h) - m(X_i,h) f_X(X_i)\right|\right|^2\right] \leq C\left(\frac{1}{nb_n^{d_x}} + b_n^{2s}\right),$$

for some constant C *that depends on* $K(\cdot)$*,* $\rho(\cdot)$*,* M_0*,* P *and* d*.*

Lemma 3.8. *Let* $r_n(z)$ *stand for* $\rho(z,\Pi_n h_0)$ *or* $\rho(z,\Pi_n h_0)\rho(z,\Pi_n h_0)^t$ *and* $m_n : x \mapsto \mathbb{E}\left[r_n(Z) \mid X = x\right]$*. Suppose that* $nb_n^{d_x} \to +\infty$*,* $\max\{\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}}, b_n^s\} = o(1)$ *and Assumptions 3.1, 3.2(ii), 3.3 and 3.4 hold. Then*

$$\max_{1\leq i\leq n} \left\|\frac{1}{nb_n^{d_x}}\sum_{j=1} K_{ij} r_n(Z_j) - m_n(X_i) f_X(X_i)\right\| = O_P\left(\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}} + b_n^s\right).$$

Lemma 3.9. *Let* Z_h *denote the vector of arguments of the functions in* $\mathcal{H}$ *and let* $\mathcal{H} := \left\{h \in L_2(P) : \sum_{l:|l|\in\{0,\ldots,m\}} \|\nabla^l h\|^2_{L_2(leb)} < +\infty\right\}$*. If* $m - d_{z_2}/2 > 0$*,* $\|\langle Z_h\rangle^\gamma\|_{L_2(P)} < +\infty$ *for some* $\gamma > 0$ *and* P^{Z_h} *has a bounded Lebesgue density* $f_{Z_h}(\cdot)$*, then for every* $M_0 > 0$*,* $\overline{\mathcal{H}}^{M_0}$ *(i.e the closure of* $\mathcal{H}^{M_0}$ *in* $L_2(P)$*) is compact in* $L_2(P)$*.*

Lemma 3.10. *Let* $f_n : z \mapsto f_n(z)$ *be some real-valued function. Suppose that* $nb_n^{d_x} \to +\infty$ *and Assumptions 3.1 and 3.4 (i)-(ii) hold. If* $\sup_{n\geq 1}\sup_{x\in[0,1]^{d_x}} \mathbb{E}[|f_n(Z)|^p \mid X = x] < +\infty$ *and* $\sup_{n\geq 1}\mathbb{E}[\max_{1\leq i\leq n}|f_n(Z_i)|^p] < +\infty$ *for some* $p \geq 2$*, then there exists* $N \geq 1$ *such that for every* $n > N$

$$\mathbb{E}\left[\sup_{x\in[0,1]^{d_x}}\left|\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n} K\left(x, X_j, b_n\right) f_n(Z_j) - \mathbb{E}\left(\frac{1}{b_n^{d_x}} K\left(x, X, b_n\right) f_n(Z)\right)\right|\right] \leq C\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}},$$

for some constant C *that depends on* d_x*,* $K(\cdot)$*,* P *and* $(f_n)_{n\geq 1}$*. The result remains valid if* $K(\cdot)$ *is replaced with* $|K(\cdot)|$*.*

Lemma 3.11. *Let* $f(z) = \sup_{h\in\mathcal{H}^{M_0}} \|\rho(z,h)\|^q$ *for some* $q \leq 2$*. Suppose that* $\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}} = o(1)$ *and Assumptions 3.1, 3.2(ii) and 3.4 (i)-(ii) hold. Then there exists* $N \geq 1$ *such that for every* $n > N$ *and every* $M_0 > 0$

$$\mathbb{E}\left[\max_{1\leq i\leq n} \frac{1}{nb_n^{d_x}}\sum_{j=1}^{n} |K_{ij}|\, f(Z_j)\right] \leq C,$$

for some constant C *that depends on* d_x*,* $K(\cdot)$*,* P *and* f*.*

Lemma 3.12. *Let* $r(z,h)$ *stand for either* $\rho(z,h)$ *or* $\rho(z,h)\rho(z,h)^t$*. Suppose that* $nb_n^{d_x} \to +\infty$*, the kernel function is bounded with support* $[-1,1]^{d_x}$ *and Assumptions 3.1, 3.2(ii), 3.4(i) and 3.6 hold. Then, there exists* $N \geq 1$ *such that for every* $n > N$ *and every* $M_0 > 0$

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\sup_{h\in\mathcal{H}^{M_0}}\left|\left|\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n} K_{ij} r(Z_j,h) - \mathbb{E}\left(\frac{1}{b_n^{d_x}} K\left(X_i, X, b_n\right) r(Z,h) \mid X_i\right)\right|\right|^2\right]\\ \leq &\frac{C}{nb_n^{d_x}},\end{aligned}$$

for some constant C *that depends on* $K(\cdot)$*,* $r(\cdot,\cdot)$*,* M_0*,* P *and* d*.*

Lemma 3.13. *Under Assumption 3.7, we have for every $M_0 > 0$ and every $(l, l') \in \{1, ..., d\}^2$*

$$\sup_{x_1 \in [0,1]^{d_x}} J_l(n, x_1, M_0, P) < +\infty \text{ and } \sup_{x_1 \in [0,1]^{d_x}} J_{l,l'}(n, x_1, M_0, P) < +\infty.$$

Lemma 3.14. *Let Assumptions 3.2(ii), 3.3(i) and (iii), and 3.5(ii) and (iii) hold. For every $M_0 > 0$, there exist positive constants $C_1 > 1$, C_2 and C_3 that depend on M_0, d_{z_h}, m and $L(\cdot)$ such that for every $n \geq 1$*

$$\begin{aligned}&\mathbb{P}\left(\sup_{h \in \mathcal{H}^{M_0} : \mathbb{E}\left[\|m(X,h)\|^2\right] \geq C_1 n^{-2m/(2m+d_{z_h})}} \left| \frac{\frac{1}{n}\sum_{i=1}^n \|m(X_i,h)\|^2}{\mathbb{E}\left[\|m(X,h)\|^2\right]} - 1 \right| \geq 0.5 \right) \\ \leq &\frac{C_2}{n^{d_{z_h}/(d_{z_h}+2m)}} e^{-C_3 n^{d_{z_h}/(d_{z_h}+2m)}}.\end{aligned}$$

3.6.2 Proofs

3.6.2.1 Proof of Lemma 3.2

Remark that for every $M_0 > 0$

$$\max_{1 \leq j \leq n} \sup_{(\lambda,h) \in \Lambda_n \times \mathcal{H}^{M_0}} \left|\lambda^t \rho(Z_j, h)\right| \leq C\mu_n \max_{1 \leq j \leq n} \sup_{h \in \mathcal{H}^{M_0}} \|\rho(Z_j, h)\| = o_{a.s}(1) = o_P(1),$$

where to get the first inequality we have used the Cauchy-Schwarz inequality and the definition of Λ_n, and to get the equality we have used $\mu_n \varphi^{-1}(n) = O(1)$ and Lemma 3.1, which is valid under Assumptions (3.1) (i) and 3.9. Let

$$\mathcal{A}_{1,n} := \left\{\omega : \psi(\lambda^t \rho(Z_j(\omega), h)) \text{ exists } \forall (j, \lambda, h) \in \{1, ..., n\} \times \Lambda_n \times \mathcal{H}^{M_0}\right\},$$

and let $\mathcal{V}_\psi$ be the domain of ψ.

Since $\mathcal{V}_\psi$ is an open interval that contains 0 and $\max_{1 \leq j \leq n} \sup_{(\lambda,h) \in \Lambda_n \times \mathcal{H}^{M_0}} |\lambda^t \rho(Z_j, h)| = o_P(1)$, there exists for every $\epsilon > 0$ an integer $N \geq 1$ such that for every $n > N$, $\mathbb{P}(\mathcal{A}_{1,n}) \geq 1 - \epsilon$.

3.6.2.2 Proof of Lemma 3.3

By definition of $\widehat{h}_n$

$$\widehat{\mathcal{L}}_n(\widehat{h}_n) + \alpha_n Pen(\widehat{h}_n) \leq \widehat{\mathcal{L}}_n(\Pi_n h_0) + \alpha_n Pen(\Pi_n h_0) + R_n.$$

What is more for every $h \in \mathcal{H}$, $\mathbf{0}_d \in \Lambda_n(h)$, so that

$$\begin{aligned}\widehat{\mathcal{L}}_n(h) &= \frac{1}{n}\sum_{i=1}^n \sup_{\lambda \in \Lambda_n(h)} \frac{1}{n b_n^{d_x}} \sum_{j=1}^n K_{ij} \psi(\lambda^t \rho(Z_j, h)) \\ &\geq \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n K_{ij} \psi(\mathbf{0}_d^t \rho(Z_j, h)) = \frac{\psi(0)}{n} \sum_{i=1}^n \frac{1}{n b_n^{d_x}} \sum_{j=1}^n K_{ij}.\end{aligned}$$

Combining the two previous inequalities, we get that for every $\epsilon > 0$ there exists $N_1 > 0$ such that for every $n > N_1$

$$1 - \frac{\epsilon}{2} \leq \mathbb{P}\Big(\frac{\psi(0)}{n} \sum_{i=1}^n \frac{1}{n b_n^{d_x}} \sum_{j=1}^n K_{ij} + \alpha_n Pen(\widehat{h}_n) \leq \widehat{\mathcal{L}}_n(\Pi_n h_0) + \alpha_n Pen(\Pi_n h_0) + R_n\Big). \tag{3.26}$$

Let $\beta_n := \sqrt{\frac{|\log b_n|}{nb_n^{d_x}}} + b_n^s$. Lemma 3.5 yields that for every $\epsilon > 0$, there exist K_1 and $N_2 \geq N_1$ such that for every $n > N_2$

$$\mathbb{P}\left(\underbrace{\widehat{\mathcal{L}}_n(\Pi_n h_0) > K_1(\beta_n^2 + \|\Pi_n h_0 - h_0\|_{\infty,\gamma}^2) + \frac{\psi(0)}{n}\sum_{i=1}^n \frac{1}{nb_n^{d_x}}\sum_{j=1}^n K_{ij}}_{=:\mathcal{A}_1} \right) < \frac{\epsilon}{2}. \tag{3.27}$$

Moreover, the discussion after Assumption 3.3 implies that

$$Pen\left(\Pi_n h_0\right) \leq M_1, \tag{3.28}$$

for some $M_1 > 0$.

With (3.27)-(3.28), we can show that for every $n > N_2$

$$\begin{aligned}
&\mathbb{P}\left(\frac{\psi(0)}{n}\sum_{i=1}^n \frac{1}{nb_n^{d_x}}\sum_{j=1}^n K_{ij} + \alpha_n Pen(\widehat{h}_n) \leq \widehat{\mathcal{L}}_n(\Pi_n h_0) + \alpha_n Pen\left(\Pi_n h_0\right) + R_n \right) \\
\leq &\mathbb{P}\left(\frac{\psi(0)}{n}\sum_{i=1}^n \frac{1}{nb_n^{d_x}}\sum_{j=1}^n K_{ij} + \alpha_n Pen(\widehat{h}_n) \leq \widehat{\mathcal{L}}_n(\Pi_n h_0) + \alpha_n M_1 + R_n \right) \\
\leq &\mathbb{P}\left(\left\{ \frac{\psi(0)}{n}\sum_{i=1}^n \frac{1}{nb_n^{d_x}}\sum_{j=1}^n K_{ij} + \alpha_n Pen(\widehat{h}_n) \leq \widehat{\mathcal{L}}_n(\Pi_n h_0) + \alpha_n M_1 + R_n \right\} \cap \mathcal{A}_1^c \right) + \mathbb{P}\left(\mathcal{A}_1\right) \\
\leq &\mathbb{P}\left(\alpha_n Pen(\widehat{h}_n) \leq 3\max\left\{K_1, M_1, 1\right\}\left\{\nu_n + \alpha_n + R_n\right\} \right) + \frac{\epsilon}{2},
\end{aligned} \tag{3.29}$$

where $\nu_n = \beta_n^2 + \|\Pi_n h_0 - h_0\|_{\infty,\gamma}^2$. Combine now (3.26) and (3.29) to obtain that for every $\epsilon > 0$, there exists $M = 3\max\left\{K_1, M_1, 1\right\}$ and $N := N_2$, such that for every $n > N$

$$\mathbb{P}\left(\alpha_n Pen(\widehat{h}_n) \leq M\left\{\nu_n + \alpha_n + R_n\right\} \right) \geq 1 - \epsilon.$$

By assumption, $(\nu_n + R_n)/\alpha_n = O(1)$, *i.e* $\sup_n \frac{\nu_n + R_n}{\alpha_n} \leq B_2$ for some $B_2 > 0$. As a result, we can say that for every $\epsilon > 0$, there exists $\widetilde{M} = M(B_2 + 1)$ such that for every $n > N$

$$\mathbb{P}\left(Pen(\widehat{h}_n) \leq \widetilde{M} \right) \geq 1 - \epsilon.$$

This is equivalent to boundedness of $Pen(\widehat{h}_n)$ in probability.

3.6.2.3 Proof of Lemma 3.4

Throughout this proof, let $\beta_n := \max\left\{\frac{1}{\sqrt{nb_n^{d_x}}}, b_n^s\right\}$. We recall that

$$\mathcal{L}_n^{\mu_n}(h) = \frac{\mu_n}{2n}\sum_{i=1}^n \|\widehat{D}(X_i, h)\|^2 - \frac{\mu_n^2}{2n}\sum_{i=1}^n \|\widehat{D}(X_i, h)\|^2 \left\| \frac{1}{nb_n^{d_x}}\sum_{j=1}^n K_{ij}\rho(Z_j, h)\rho(Z_j, h)^t \right\|.$$

Thanks to the inequality $\|a\|^2 \geq \frac{1}{2}\|b\|^2 - \|b-a\|^2$ as well as the Cauchy-Schwarz and triangle inequalities, we have for every $h \in \mathcal{H}_{\delta_{1,n},\delta_{2,n}}^{M_0}$

$$\begin{aligned}
\mathcal{L}_n^{\mu_n}(h) \geq &\frac{\mu_n}{2}\left\{ \frac{1}{2n}\sum_{i=1}^n \|m(X_i, h)f_X(X_i)\|^2 - \frac{1}{n}\sum_{i=1}^n \|\widehat{D}(X_i, h) - m(X_i, h)f_X(X_i)\|^2 \right. \\
&\left. - \mu_n \underbrace{\frac{1}{n}\sum_{i=1}^n \|\widehat{D}(X_i, h)\|^2 \left\| \frac{1}{nb_n^{d_x}}\sum_{j=1}^n K_{ij}\rho(Z_j, h)\rho(Z_j, h)^t \right\|}_{:=A_2} \right\}.
\end{aligned}$$

Taking the infimum over $\mathcal{H}^{M_0}_{\delta_{1,n},\delta_{2,n}}$ on both sides

$$\inf_{h\in\mathcal{H}^{M_0}_{\delta_{1,n},\delta_{2,n}}} \mathcal{L}_n^{\mu_n}(h) \geq \frac{\mu_n}{2} \inf_{h\in\mathcal{H}^{M_0}_{\delta_{1,n},\delta_{2,n}}} \left\{ \frac{1}{2n}\sum_{i=1}^n \|m(X_i,h)f_X(X_i)\|^2 \right.$$
$$-\frac{1}{n}\sum_{i=1}^n \|\widehat{D}(X_i,h) - m(X_i,h)f_X(X_i)\|^2$$
$$\left. -\mu_n \underbrace{\frac{1}{n}\sum_{i=1}^n \|\widehat{D}(X_i,h)\|^2 \left\| \frac{1}{nb_n^{d_x}} \sum_{j=1}^n K_{ij}\rho(Z_j,h)\rho(Z_j,h)^t \right\|}_{:=A_2} \right\}. \tag{3.30}$$

By Lemma 3.7 and Markov's inequality, we know that there exists $N_1 \geq 1$ (independent of ϵ) such that for every $n > N_1$

$$\mathbb{P}\left(\underbrace{\frac{1}{n}\sum_{i=1}^n \sup_{h\in\mathcal{H}^{M_0}} \|\widehat{D}(X_i,h) - m(X_i,h)f_X(X_i)\|^2 \leq \frac{C_1}{\epsilon}\beta_n^2}_{:=\mathcal{A}_{1,n}} \right) \geq 1 - \frac{\epsilon}{6}, \tag{3.31}$$

where C_1 is a constant that depends neither on n nor ϵ.

We now construct an upper bound on A_2. Let $\mathbb{V}(\cdot,h) = \mathbb{E}[\rho(Z,h)\rho(Z,h)^t \mid X = \cdot]$. The triangle inequality first yields uniformly over $\mathcal{H}^{M_0}_{\delta_{1,n},\delta_{2,n}}$

$$\begin{aligned} A_2 \leq & \frac{1}{n}\sum_{i=1}^n \|\widehat{D}(X_i,h)\|^2 \times \left\{ \sup_{h\in\mathcal{H}^{M_0}} \left\| \frac{1}{nb_n^{d_x}} \sum_{j=1}^n K_{ij}\rho(Z_j,h)\rho(Z_j,h)^t - \mathbb{V}(X_i,h)\, f_X(X_i) \right\| \right. \\ & \left. + \sup_{h\in\mathcal{H}^{M_0}} ||\mathbb{V}(X_i,h)\, f_X(X_i)|| \right\} \\ \leq & \frac{1}{n}\sum_{i=1}^n \|\widehat{D}(X_i,h)\|^2 \times \left\{ + \sup_{h\in\mathcal{H}^{M_0}} \left\| \frac{1}{nb_n^{d_x}} \sum_{j=1}^n K_{ij}\rho(Z_j,h)\rho(Z_j,h)^t - \mathbb{V}(X_i,h)\, f_X(X_i) \right\| \right. \\ & \left. + C_2 \right\}, \end{aligned}$$

where $C_2 = \overline{f}_X \sup_{(x,h)\in[0,1]^{d_x}\times\mathcal{H}^{M_0}} \|\mathbb{V}(x,h)\|$.

The inequality $(|a| + |b|)^2 \leq 2(a^2 + b^2)$ and repeated use of the Cauchy-Schwarz inequality imply

uniformly over $\mathcal{H}^{M_0}_{\delta_{1,n},\delta_{2,n}}$

$$\begin{aligned}A_2 \leq & \frac{2C_2}{n}\sum_{i=1}^{n}\|m(X_i,h)f_X(X_i)\|^2 + \frac{2C_2}{n}\sum_{i=1}^{n}\sup_{h\in\mathcal{H}^{M_0}}\|\widehat{D}(X_i,h)-m(X_i,h)f_X(X_i)\|^2 \\ & + 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|m(X_i,h)f_X(X_i)\|^4} \\ & \times\sqrt{\frac{1}{n}\sum_{i=1}^{n}\sup_{h\in\mathcal{H}^{M_0}}\left\|\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K_{ij}\rho(Z_j,h)\rho(Z_j,h)^t-\mathbb{V}\left(X_i,h\right)f_X(X_i)\right\|^2} \\ & + 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\sup_{h\in\mathcal{H}^{M_0}}\|\widehat{D}(X_i,h)-m(X_i,h)f_X(X_i)\|^4} \\ & \times\sqrt{\frac{1}{n}\sum_{i=1}^{n}\sup_{h\in\mathcal{H}^{M_0}}\left\|\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K_{ij}\rho(Z_j,h)\rho(Z_j,h)^t-\mathbb{V}\left(X_i,h\right)f_X(X_i)\right\|^2}, \end{aligned} \tag{3.32}$$

By Lemma 3.12, we can claim that there exists $N_2 \geq N_1$ such that for every $n > N_2$, every $\epsilon > 0$ and a constant C_3 that does not depend on ϵ and n

$$1-\frac{\epsilon}{6} \leq \mathbb{P}\left(\underbrace{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\sup_{h\in\mathcal{H}^{M_0}}\left\|\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K_{ij}\rho(Z_j,h)\rho(Z_j,h)^t-\mathbb{V}\left(X_i,h\right)f_X(X_i)\right\|^2} \leq \frac{C_3\beta_n}{\sqrt{\epsilon}}}_{:=\mathcal{A}_{2,n}}\right). \tag{3.33}$$

We note that

$$\begin{aligned} & \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sup_{h\in\mathcal{H}^{M_0}}\left\|\widehat{D}(X_i,h)-m\left(X_i,h\right)f_X(X_i)\right\|^4} \\ \leq & \sqrt{2}\sqrt{\max_{1\leq i\leq n}\sup_{h\in\mathcal{H}^{M_0}}\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}|K_{ij}|\,\|\rho(Z_j,h)\|^2+C_4^2} \\ & \times\sqrt{\frac{1}{n}\sum_{i=1}^{n}\sup_{h\in\mathcal{H}^{M_0}}\left\|\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K_{ij}\rho(Z_j,h)-m\left(X_i,h\right)f_X(X_i)\right\|^2}, \end{aligned}$$

where $C_4 := \sup_{(x,h)\in[0,1]^{d_x}\times\mathcal{H}^{M_0}}\|m(x,h)\|\,\overline{f}_X$. By Lemmas 3.11 (with $p=2$) and 3.12, and (3.31), we can claim there exists $N_3 \geq N_2$ such that for every $n > N_3$, every $\epsilon > 0$ and a constant C_5 that does not depend on ϵ and n

$$1-\frac{\epsilon}{6} \leq \mathbb{P}\left(\underbrace{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\sup_{h\in\mathcal{H}^{M_0}}\|\widehat{D}(X_i,h)-m(X_i,h)f_X(X_i)\|^4} \leq \sqrt{2}\sqrt{\frac{C_5}{\epsilon}+C_4^2}\sqrt{\frac{C_1}{\epsilon}}\beta_n}_{:=\mathcal{A}_{3,n}}\right). \tag{3.34}$$

Let $A_3 := \frac{1}{n}\sum_{i=1}^{n}\|m(X_i,h)\|^2$ and

$$A_4 := C_6A_5 + \frac{C_7}{\sqrt{\epsilon}}\beta_n\sup_{h\in\mathcal{H}^{M_0}_{\delta_{1,n},\delta_{2,n}}}\sqrt{A_5} + C_8\max\left\{\epsilon^{-3/2},\epsilon^{-1}\right\}\beta_n^2,$$

where $C_6 := 2C_2\overline{f}_X^2$, $C_7 := 2\overline{f}_XC_3C_4$ and $C_8 := 4\sqrt{2}\sqrt{C_1}C_3\max\left\{\sqrt{C_5},C_4\right\}$.

We can now gather (3.30), (3.31), (3.32), (3.33) and (3.34) and use the fact that $\underline{f}_X \leq f_X(X_i) \leq \overline{f}_X$ to claim that for every $n > N_3$ and every $\epsilon > 0$

$$\begin{aligned}1 - \frac{\epsilon}{2} \leq & \mathbb{P}\left(\mathcal{A}_{1,n} \cap \mathcal{A}_{2,n} \cap \mathcal{A}_{3,n}\right) \\ \leq & \mathbb{P}\left(\inf_{h \in \mathcal{H}^{M_0}_{\delta_{1,n},\delta_{2,n}}} \mathcal{L}_n^{\mu_n}(h) \geq \frac{\mu_n}{2} \inf_{h \in \mathcal{H}^{M_0}_{\delta_{1,n},\delta_{2,n}}} \left\{\frac{1}{4}\underline{f}_X^2 A_5 - \frac{C_1 \beta_n^2}{\epsilon} - \mu_n A_4\right\}\right).\end{aligned}$$

Since $\mu_n = o(1)$, we can claim that there exists $N_4 \geq N_3$ such that for every $n > N_4$, every $\epsilon > 0$ and some C_9 independent of n and ϵ

$$\begin{aligned}&\mathbb{P}\left(\inf_{h \in \mathcal{H}^{M_0}_{\delta_{1,n},\delta_{2,n}}} \mathcal{L}_n^{\mu_n}(h)\right. \\ &\qquad \left.\geq \mu_n \inf_{h \in \mathcal{H}^{M_0}_{\delta_{1,n},\delta_{2,n}}} \left\{\frac{1}{16}\underline{f}_X^2 A_5 - C_9 \beta_n^2 \max\left\{\epsilon^{-3/2}, \epsilon^{-1}\right\} - \mu_n \frac{C_9}{\sqrt{\epsilon}} \beta_n \sup_{h \in \mathcal{H}^{M_0}_{\delta_{1,n},\delta_{2,n}}} \sqrt{A_5}\right\}\right) \\ &\geq 1 - \frac{\epsilon}{2}. \end{aligned} \tag{3.35}$$

Let C_1 be as in the statement of Lemma 3.14. We impose $\delta_{1,n} \geq C_1 n^{-2m/(2m+d_{z_h})}$ so that Lemma 3.14 holds and allows us to conclude that for every $n \geq 1$

$$\begin{aligned}&\mathbb{P}\left(\forall h \in \mathcal{H}^{M_0}_{\delta_{1,n},\delta_{2,n}} : 0.5\mathbb{E}\left[\|m(X,h)\|^2\right] \leq A_6 \leq 2\mathbb{E}\left[\|m(X,h)\|^2\right]\right) \\ \geq & 1 - \frac{C_{10}}{n^{d_{z_h}/(d_{z_h}+2m)}} e^{-C_{11} n^{d_{z_h}/(d_{z_h}+2m)}}.\end{aligned} \tag{3.36}$$

Let $A_6 := C_9 \beta_n^2 \max\left\{\epsilon^{-3/2}, \epsilon^{-1}\right\} + 2\mu_n \frac{C_9}{\sqrt{\epsilon}} \beta_n \sup_{h \in \mathcal{H}^{M_0}_{\delta_{1,n},\delta_{2,n}}} \sqrt{\mathbb{E}\left[\|m(X,h)\|^2\right]}$. Combining (3.35) and (3.36), we obtain that for every $n > N_4$ and every $\epsilon > 0$

$$\begin{aligned}&\mathbb{P}\left(\inf_{h \in \mathcal{H}^{M_0}_{\delta_{1,n},\delta_{2,n}}} \mathcal{L}_n^{\mu_n}(h) \geq \mu_n \left(\frac{\underline{f}_X^2}{32} \inf_{h \in \mathcal{H}^{M_0}_{\delta_{1,n},\delta_{2,n}}} \mathbb{E}\left[\|m(X,h)\|^2\right] - A_6\right)\right) \\ \geq & 1 - \frac{\epsilon}{2} - \frac{C_{10}}{n^{d_{z_h}/(d_{z_h}+2m)}} e^{-C_{11} n^{d_{z_h}/(d_{z_h}+2m)}}.\end{aligned}$$

3.6.2.4 Proof of Lemma 3.5

From the discussion following Assumption 3.3, we know that there exists some $M_0 \in \mathbb{R}_+^*$ such that for every $n \geq 1$, $\Pi_n h_0 \in \mathcal{H}^{M_0}$ w.p.1. Assumption 3.2 implies that Lemma 3.2 is applicable. This yields that w.p.a.1., $\psi\left(\lambda^t \rho(Z_j, \Pi_n h_0)\right)$ is well-defined for every $1 \leq j \leq n$ uniformly in $\lambda \in \Lambda_n$ where $\Lambda_n := \left\{\lambda \in \mathbb{R}^d : \|\lambda\| \leq n^{-1/p}\right\}$.

By continuity of the function $\xi \mapsto \psi(\xi)$ and compactness of Λ_n, for every $i \in \{1, \ldots, n\}$, $\sup_{\lambda \in \Lambda_n} \frac{1}{n b_n^{d_x}} \sum_{j=1}^n K_{ij} \psi\left(\lambda^t \rho(Z_j, \Pi_n h_0)\right)$ is attained at some $\lambda \in \Lambda_n$ that we denote $\widehat{\lambda}_i$. By a Mean Value expansion of $\psi(\xi)$ around $\xi = 0$ and the fact that $\psi'(0) = -1$, for every $i \in \{1, ..., n\}$, there exists $\tau \in (0,1)$ such that

$$\begin{aligned}&\frac{1}{n b_n^{d_x}} \sum_{j=1}^n K_{ij} \psi\left(\widehat{\lambda}_i^t \rho(Z_j, \Pi_n h_0)\right) \\ = & \psi(0) \frac{1}{n b_n^{d_x}} \sum_{j=1}^n K_{ij} - \widehat{\lambda}_i^t \widehat{D}(X_i, \Pi_n h_0) + \frac{1}{2 n b_n^{d_x}} \sum_{j=1}^n K_{ij} (\widehat{\lambda}_i^t \rho(Z_j, \Pi_n h_0))^2 \psi''(\tau \widehat{\lambda}_i^t \rho(Z_j, \Pi_n h_0)).\end{aligned} \tag{3.37}$$

What is more, $\mathbf{0}_d \in \Lambda_n$ by construction which implies

$$\begin{aligned}\psi(0)\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K_{ij} =& \frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K_{ij}\psi\left(\mathbf{0}_d^t\rho(Z_j,\Pi_n h_0)\right)\\ \leq& \frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K_{ij}\psi\left(\widehat{\lambda}_i^t\rho(Z_j,\Pi_n h_0)\right).\end{aligned} \tag{3.38}$$

Let $V_n : x \mapsto \mathbb{E}\left[\rho(Z,\Pi_n h_0)\rho(Z,\Pi_n h_0)^t \mid X = x\right]$. Combining (3.37), (3.38) and the Cauchy-Schwarz and triangle inequalities, we get for every $i \in \{1,...,n\}$

$$\begin{aligned}0 \leq& -\widehat{\lambda}_i^t\widehat{D}(X_i,\Pi_n h_0) + \frac{1}{2nb_n^{d_x}}\sum_{j=1}^{n}K_{ij}(\widehat{\lambda}_i^t\rho(Z_j,\Pi_n h_0))^2\psi^{''}(\tau\widehat{\lambda}_i^t\rho(Z_j,\Pi_n h_0))\\
\leq& \left\|\widehat{\lambda}_i\right\|\left\|\widehat{D}(X_i,\Pi_n h_0)\right\| + \frac{1}{2}\left\{\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K_{ij}\left(\psi^{''}(\tau\widehat{\lambda}_i^t\rho(Z_j,\Pi_n h_0))+1\right)(\widehat{\lambda}_i^t\rho(Z_j,\Pi_n h_0))^2\right.\\
&-\widehat{\lambda}_i^t\left(\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K_{ij}\rho(Z_j,\Pi_n h_0)\rho(Z_j,\Pi_n h_0)^t - V_n(X_i)f_X(X_i)\right)\widehat{\lambda}_i\\
&\left.-\widehat{\lambda}_i^t V_n(X_i)f_X(X_i)\widehat{\lambda}_i\right\}\\
\leq& \left\|\widehat{\lambda}_i\right\|\left\|\widehat{D}(X_i,\Pi_n h_0)\right\| + \frac{1}{2}\left|\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K_{ij}\left(\psi^{''}(\tau\widehat{\lambda}_i^t\rho(Z_j,\Pi_n h_0))+1\right)(\widehat{\lambda}_i^t\rho(Z_j,\Pi_n h_0))^2\right|\\
&+\frac{1}{2}\left\|\widehat{\lambda}_i\right\|^2\left\|\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K_{ij}\rho(Z_j,\Pi_n h_0)\rho(Z_j,\Pi_n h_0)^t - V_n(X_i)f_X(X_i)\right\|\\
&-\frac{C}{2}\left\|\widehat{\lambda}_i\right\|^2 = \left\|\widehat{\lambda}_i\right\|\left\|\widehat{D}(X_i,\Pi_n h_0)\right\| + A_1 + A_2 - \frac{C}{2}\left\|\widehat{\lambda}_i\right\|^2,\end{aligned} \tag{3.39}$$

where on the last line, we used Assumption 3.2(i) and Assumption 3.4(i).

We first control A_1. We remark

$$\max_{1\leq i,j\leq n}\left|\psi^{''}(\tau\widehat{\lambda}_i^t\rho(Z_j,\Pi_n h_0))+1\right| \leq \max_{1\leq j\leq n}\sup_{(\tau,v,h)\in[0,1]\times\Lambda_n\times\mathcal{H}^{M_0}}\left|\psi^{''}(\tau v^t\rho(Z_j,h))+1\right|.$$

For every $\psi(\cdot)$, we know there exists a compact interval I that strictly includes 0 and over which $\psi^{''}(\cdot)$ is Lipschitz. Under Assumption 3.2, Lemma 3.1 is applicable. Thanks to this lemma and the definition of Λ_n, we can claim that $\tau v^t\rho(Z_j,h)$ belongs to I ***w.p.a.1*** uniformly in $(j,\tau,v,h) \in \{1,...,n\}\times[0,1]\times\Lambda_n\times\mathcal{H}^{M_0}$ and $\max_{1\leq j\leq n}\sup_{(\tau,v,h)\in[0,1]\times\Lambda_n\times\mathcal{H}^{M_0}}|\tau v^t\rho(Z_j,h)| = o_P(1)$. As result $\max_{1\leq i,j\leq n}\left|\psi^{''}(\tau\widehat{\lambda}_i^t\rho(Z_j,\Pi_n h_0))+1\right| = o_P(1)$ by continuous mapping and

$$\begin{aligned}A_1 \leq& \frac{1}{2}\left\|\widehat{\lambda}_i\right\|^2\max_{1\leq j\leq n}\left|\psi^{''}(\tau\widehat{\lambda}_i^t\rho(Z_j,\Pi_n h_0))+1\right|\max_{1\leq i\leq n}\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}|K_{ij}|\left\|\rho(Z_j,\Pi_n h_0)\right\|^2\\
=& \left\|\widehat{\lambda}_i\right\|^2 o_P\left(\max_{1\leq i\leq n}\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}|K_{ij}|\left\|\rho(Z_j,\Pi_n h_0)\right\|^2\right).\end{aligned}$$

We observe that

$$\max_{1\leq i\leq n}\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}|K_{ij}|\left\|\rho(Z_j,\Pi_n h_0)\right\|^2 \leq \max_{1\leq i\leq n}\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}|K_{ij}|\sup_{h\in\mathcal{H}^{M_0}}\left\|\rho(Z_j,h)\right\|^2 = O_P(1)$$

by Lemma 3.11 with $f(z) = \sup_{h\in\mathcal{H}^{M_0}}\left\|\rho(z,h)\right\|^2$ and Markov's inequality. We can thus write

$$A_1 \leq \left\|\widehat{\lambda}_i\right\|^2 o_P(1),$$

where the $o_P(1)$ term is uniform in $i \in \{1, ..., n\}$.

The term A_2 can be controlled thanks to Lemma 3.8

$$\begin{aligned} A_2 \leq & \frac{1}{2} \left\| \widehat{\lambda}_i \right\|^2 \max_{1 \leq i \leq n} \left\| \frac{1}{n b_n^{d_x}} \sum_{j=1}^{n} K_{ij} \rho(Z_j, \Pi_n h_0) \rho(Z_j, \Pi_n h_0)^t - V_n(X_i) f_X(X_i) \right\| \\ = & \left\| \widehat{\lambda}_i \right\|^2 O_P \left(\sqrt{\frac{|\log b_n|}{n b_n^{d_x}}} + b_n^s \right) = \left\| \widehat{\lambda}_i \right\|^2 o_P(1), \end{aligned}$$

where once again the $o_P(1)$ term is uniform in $i \in \{1, ..., n\}$.

Based on (3.39), we can therefore claim that *w.p.a.1* uniformly in $i \in \{1, ..., n\}$,

$$\begin{aligned} 0 \leq & \left\| \widehat{\lambda}_i \right\| \left\| \widehat{D}(X_i, \Pi_n h_0) \right\| - \frac{C}{4} \left\| \widehat{\lambda}_i \right\|^2 \\ \implies & \left\| \widehat{\lambda}_i \right\| \leq \frac{4}{C} \max_{1 \leq i \leq n} \left\| \widehat{D}(X_i, \Pi_n h_0) \right\|. \end{aligned}$$

Lemma 3.8, Assumptions 3.4(i) and 3.5(iii) and $\max \left\{ \sqrt{\frac{|\log b_n|}{n b_n^{d_x}}}, b_n^s, \|\Pi_n h_0 - h_0\|_{\infty, \gamma} \right\} = o(n^{-1/p})$ induce that *w.p.a.1* uniformly in $i \in \{1, ..., n\}$,

$$\left\| \widehat{\lambda}_i \right\| \leq O_P \left(\sqrt{\frac{|\log b_n|}{n b_n^{d_x}}} + b_n^s + \|\Pi_n h_0 - h_0\|_{\infty, \gamma} \right) = o_P \left(n^{-1/p} \right).$$

Following the final steps of [113, Lemma A2], we can conclude that *w.p.a.1* uniformly in $i \in \{1, ..., n\}$, $\widehat{\lambda}_i = \operatorname{argmax}_{\lambda \in \Lambda_n(\Pi_n h_0)} \frac{1}{n b_n^{d_x}} \sum_{j=1}^{n} K_{ij} \psi \left(\lambda^t \rho(Z_j, \Pi_n h_0) \right)$ and

$$\begin{aligned} & \frac{1}{n b_n^{d_x}} \sum_{j=1}^{n} K_{ij} \psi \left(\widehat{\lambda}_i^t \rho(Z_j, \Pi_n h_0) \right) = \sup_{\lambda \in \Lambda_n(\Pi_n h_0)} \frac{1}{n b_n^{d_x}} \sum_{j=1}^{n} K_{ij} \psi \left(\lambda^t \rho(Z_j, \Pi_n h_0) \right) \\ \leq & \psi(0) \frac{1}{n b_n^{d_x}} \sum_{j=1}^{n} K_{ij} + O_P \left(\frac{|\log b_n|}{n b_n^{d_x}} + b_n^{2s} + \|\Pi_n h_0 - h_0\|_{\infty, \gamma}^2 \right). \end{aligned}$$

3.6.2.5 Proof of Lemma 3.6

Lemma 3.1 is valid so that for every $M_0 > 0$

$$\max_{1 \leq i \leq n} \sup_{h \in \mathcal{H}^{M_0}} \|\rho(Z_i, h)\| = o_{a.s} \left(\varphi^{-1}(n) \right).$$

This implies that for every positive M_0, ϵ and δ, there exists $N \geq 1$ such that for every $n > N$

$$\mathbb{P} \left(\max_{1 \leq i \leq n} \sup_{h \in \mathcal{H}^{M_0}} \|\rho(Z_i, h)\| \leq \varphi^{-1}(n) \delta \right) \geq 1 - \epsilon.$$

When $\max_{1 \leq i \leq n} \sup_{h \in \mathcal{H}^{M_0}} \|\rho(Z_i, h)\| \leq \varphi^{-1}(n) \delta$, we have

$$\max_{1 \leq i \leq n} \sup_{(\tau, v, h) \in [0,1] \times \Lambda_n \times \mathcal{H}^{M_0}} \left| \tau v^t \rho(Z_i, h) \right| \leq \mu_n C \max_{1 \leq i \leq n} \|\rho(Z_i, h)\| \leq C_1 \delta,$$

where $C_1 := C \times \sup_{n \geq 1} \mu_n \varphi^{-1}(n)$ is finite by assumption.

By construction, for every $\psi(\cdot)$ function we consider there exists a compact interval I such that 0 is a strict subset of I and $\psi''(\cdot)$ is Lipschitz over I. The $\psi(\cdot)$ functions we are interested in also satisfy $\psi''(0) = -1$. As a result, for every C and $\psi(\cdot)$ we can find δ such that whenever $\max_{1 \leq i \leq n} \sup_{h \in \mathcal{H}^{M_0}} \|\rho(Z_i, h)\| \leq \varphi^{-1}(n) \delta$, we get

$$\max_{1 \leq i \leq n} \sup_{(\tau, v, h) \in [0,1] \times \Lambda_n \times \mathcal{H}^{M_0}} \left| \psi''(\tau v^t \rho(Z_i, h)) + 1 \right| \leq 1$$

Those findings allow us to conclude that for every $\psi(\cdot)$ in the GEL family we consider, and every positive M_0, C and ϵ, there exist N and δ such that for every $n > N$

$$\begin{aligned}1-\epsilon \leq & \mathbb{P}\left(\max_{1\leq i\leq n}\sup_{h\in\mathcal{H}^{M_0}}\|\rho(Z_i,h)\|\leq \varphi^{-1}(n)\delta\right)\\ \leq & \mathbb{P}\left(\max_{1\leq i\leq n}\sup_{(\tau,v,h)\in[0,1]\times\Lambda_n\times\mathcal{H}^{M_0}}\left|\psi^{''}(\tau v^t\rho(Z_i,h))+1\right|\leq 1\right).\end{aligned}$$

3.6.2.6 Proof of Lemma 3.7

Let $(X,Z)\sim P$ and $(X,Z)\perp\!\!\!\perp (X_i,Z_i)_{i=1}^n$. By the triangle inequality, a convexity argument and Assumption 3.1

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\sup_{h\in\mathcal{H}^{M_0}}\left\|\widehat{D}(X_i,h)-m(X_i,h)f_X(X_i)\right\|^2\right]\\ \leq & 2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\sup_{h\in\mathcal{H}^{M_0}}\left\|\widehat{D}(X_i,h)-\mathbb{E}\left(\frac{1}{b_n^{d_x}}K\left(X_i,X,b_n\right)\rho(Z,h)\mid X_i\right)\right\|^2\right]\\ &+2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\sup_{h\in\mathcal{H}^{M_0}}\left\|\mathbb{E}\left(\frac{1}{b_n^{d_x}}K\left(X_i,X,b_n\right)\rho(Z,h)\mid X_i\right)-m(X_i,h)f_X(X_i)\right\|^2\right]\\ \leq & 2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\sup_{h\in\mathcal{H}^{M_0}}\left\|\widehat{D}(X_i,h)-\mathbb{E}\left(\frac{1}{b_n^{d_x}}K\left(X_i,X,b_n\right)\rho(Z,h)\right)\right\|^2\right]\\ &+2\sup_{(x_1,h)\in[0,1]^{d_x}\times\mathcal{H}^{M_0}}\left\|\mathbb{E}\left(\frac{1}{b_n^{d_x}}K\left(x_1,X,b_n\right)\rho(Z,h)\right)-m(x_1,h)f_X(x_1)\right\|^2\\ =: & A_1+A_2.\end{aligned}$$

Thanks to Lemma 3.12, we can claim that there exists $N_1\geq 1$ such that for every $n>N_1$

$$A_1\leq\frac{C_1}{nb_n^{d_x}},\tag{3.40}$$

for some C that depends on $K(\cdot)$, $\rho(\cdot)$, M_0, P and d.

We now control term A_2. Recall that

$$I_{b_n}:=\left\{x\in[0,1]^{d_x}:\exists t\in\{1,...,d_x\}\text{ such that }x^{(t)}<b_n\text{ or }x^{(t)}>1-b_n\right\}$$

and let $I_{b_n}^c:=[0,1]^{d_x}\setminus I_{b_n}$. We first control

$$\sup_{(x_1,h)\in I_{b_n}^c\times\mathcal{H}^{M_0}}\left\|\mathbb{E}\left(\frac{1}{b_n^{d_x}}K\left(x_1,X,b_n\right)\rho(Z,h)\right)-m(x_1,h)f_X(x_1)\right\|^2.$$

Let $\mathcal{U}_{x_1,b_n}:=\left\{u\in\mathbb{R}^{d_x}:u=\frac{x_1-x}{b_n},x\in[0,1]^{d_x}\right\}$. Note that for every $x_1\in I_{b_n}^c$, $K\left(x_1,X,b_n\right)$ is actually equal to $\prod_{t=1}^{d_x}\widetilde{K}\left(\frac{x_1^{(t)}-X^{(t)}}{b_n}\right)=:K\left(x_1,X,b_n\right)$ and

$$\begin{aligned}\mathbb{E}\left[\frac{1}{b_n^{d_x}}K\left(x_1,X,b_n\right)\rho_l(Z,h)\right]=&\frac{1}{b_n^{d_x}}\int_{[0,1]^{d_x}}K\left(x_1,x,b_n\right)\mathbb{E}[\rho_l(Z,h)\mid X=x]f_X(x)dx\\ =&\int_{\mathcal{U}_{x_1,b_n}}\prod_{t=1}^{d_x}\widetilde{K}\left(u^{(t)}\right)\mathbb{E}[\rho_l(Z,h)\mid X=x_1-b_nu]f_X(x_1-b_nu)du.\end{aligned}\tag{3.41}$$

As the support of $\widetilde{K}(\cdot)$ is $[-1,1]$, there exists $N_2 \geq N_1$ such that for every $n > N_2$ and every $x_1 \in I_{b_n}^c$

$$\begin{aligned}&\int_{\mathcal{U}_{x_1,b_n}} \prod_{t=1}^{d_x} \widetilde{K}\left(u^{(t)}\right) \mathbb{E}[\rho_l(Z,h) \mid X = x_1 - b_n u] f_X(x_1 - b_n u) du \\ =&\int_{[-1,1]^{d_x}} \prod_{t=1}^{d_x} \widetilde{K}\left(u^{(t)}\right) \mathbb{E}[\rho_l(Z,h) \mid X = x_1 - b_n u] f_X(x_1 - b_n u) du.\end{aligned}$$

For every $x_1 \in I_{b_n}^c$, Assumption 3.4(i)-(ii) allows us to do a Taylor-Lagrange expansion of order $s-1$ around 0 of $u \mapsto \mathbb{E}[\rho_l(Z,h) \mid X = x_1 - b_n u] f_X(x_1 - b_n u)$: there exists a $\tau \in (0,1)$ (possibly depending on u) such that

$$\begin{aligned}&\int_{[-1,1]^{d_x}} \prod_{t=1}^{d_x} \widetilde{K}\left(u^{(t)}\right) \mathbb{E}[\rho_l(Z,h) \mid X = x_1 - b_n u] f_X(x_1 - b_n u) du \\ =&\int_{[-1,1]^{d_x}} \prod_{t=1}^{d_x} \widetilde{K}\left(u^{(t)}\right) \left\{ \sum_{\beta:|\beta| \in \{0,\ldots,s-1\}} D^\beta \left\{\mathbb{E}[\rho_l(Z,h) \mid X = x_1] f_X(x_1)\right\} b_n^{|\beta|} u^\beta \right. \\ &\left. + \sum_{\beta:|\beta|=s} D^\beta \left\{\mathbb{E}[\rho_l(Z,h) \mid X = x_1 - \tau b_n u] f_X(x_1 - \tau b_n u)\right\} b_n^s u^\beta \right\} du. \end{aligned} \tag{3.42}$$

Since the $\widetilde{K}(\cdot)$ is of order s, we have $\int_{[-1,1]^{d_x}} \prod_{t=1}^{d_x} \widetilde{K}\left(u^{(t)}\right) du = 1$ and $\int_{[-1,1]^{d_x}} u^\beta \prod_{t=1}^{d_x} \widetilde{K}\left(u^{(t)}\right) du = 0$ for every $\beta : |\beta| \in \{1,...,s-1\}$. This and (3.42) imply

$$\begin{aligned}&\int_{[-1,1]^{d_x}} \prod_{t=1}^{d_x} \widetilde{K}\left(u^{(t)}\right) \left\{ \sum_{\beta:|\beta| \in \{0,\ldots,s-1\}} D^\beta \left\{\mathbb{E}[\rho_l(Z,h) \mid X = x_1] f_X(x_1)\right\} b_n^{|\beta|} u^\beta \right. \\ &\left. + \sum_{\beta:|\beta|=s} D^\beta \left\{\mathbb{E}[\rho_l(Z,h) \mid X = x_1 - \tau b_n u] f_X(x_1 - \tau b_n u)\right\} b_n^s u^\beta \right\} du \\ =&\mathbb{E}[\rho_l(Z,h) \mid X = x_1] f_X(x_1) \\ &+ \int_{[-1,1]^{d_x}} \prod_{t=1}^{d_x} \widetilde{K}\left(u^{(t)}\right) \sum_{\beta:|\beta|=s} D^\beta \left\{\mathbb{E}[\rho_l(Z,h) \mid X = x_1 - \tau b_n u] f_X(x_1 - \tau b_n u)\right\} b_n^s u^\beta du. \end{aligned} \tag{3.43}$$

Under Assumption 3.4(i)-(ii), for every β such that $|\beta| \in \{1,...,s\}$ and for every $l \in \{1,...,d\}$, it holds

$$\sup_{(x,h) \in [0,1]^{d_x} \times \mathcal{H}^{M_0}} \left| D^\beta \left\{\mathbb{E}[\rho_l(Z,h) \mid X = x] f_X(x)\right\} \right| < +\infty. \tag{3.44}$$

We combine (3.41)-(3.44) and we use Assumption 3.4(iii) to conclude that

$$\begin{aligned}&\sup_{(x_1,h) \in I_{b_n}^c \times \mathcal{H}^{M_0}} \left\| \mathbb{E}\left(\frac{1}{b_n^{d_x}} K(x_1, X, b_n) \rho(Z,h)\right) - m(x_1,h) f_X(x_1) \right\|^2 \\ \leq& d \max_{l \in \{1,\ldots,d\}} \sup_{(x_1,h) \in I_{b_n}^c \times \mathcal{H}^{M_0}} \left| \mathbb{E}\left(\frac{1}{b_n^{d_x}} K(x_1, X, b_n) \rho_l(Z,h)\right) - \mathbb{E}\left[\rho_l(Z,h) \mid X = x_1\right] f_X(x_1) \right|^2 \\ \leq& d \max_{l \in \{1,\ldots,d\}} \sup_{(x,h) \in I_{b_n}^c \times \mathcal{H}^{M_0}} \end{aligned} \tag{3.45}$$

$$\left| \int_{[-1,1]^{d_x}} \prod_{t=1}^{d_x} \widetilde{K}\left(u^{(t)}\right) \sum_{\beta:|\beta|=s} D^\beta \left\{\mathbb{E}[\rho_l(Z,h) \mid X = x - \tau b_n u] f_X(x - \tau b_n u)\right\} b_n^s u^\beta du \right|^2$$

$$\begin{aligned}\leq& b_n^{2s} d \max_{l \in \{1,\ldots,d\}} \sup_{(x,h) \in I_{b_n}^c \times \mathcal{H}^{M_0}} \left| D^\beta \mathbb{E}[\rho_l(Z,h) \mid X = x] f_X(x) \right|^2 \left(\sum_{\beta:|\beta|=s} \int_{[-1,1]^{d_x}} \left| \prod_{t=1}^{d_x} \widetilde{K}\left(u^{(t)}\right) u^\beta \right| du \right)^2 \\ =& C_2 b_n^{2s}. \end{aligned} \tag{3.46}$$

An analogous reasoning allows us to claim

$$\sup_{(x_1,h)\in I_{b_n}\times\mathcal{H}^{M_0}}\left\|\mathbb{E}\left(\frac{1}{b_n^{d_x}}K\left(x_1,X,b_n\right)\rho(Z,h)\right)-m(x_1,h)f_X(x_1)\right\|^2\leq C_3 b_n^{2s}. \tag{3.47}$$

Finally, (3.40), (3.45) and (3.47) ensure that there exists $N\geq 1$ such that for every

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\sup_{h\in\mathcal{H}^{M_0}}\left|\left|\widehat{D}(X_i,h)-m(X_i,h)f_X(X_i)\right|\right|^2\right]\leq C_4\left(\frac{1}{nb_n^{d_x}}+b_n^{2s}\right),$$

where $C_4=2\max\{C_1,C_2,C_3\}$.

3.6.2.7 Proof of Lemma 3.8

The proof is very similar to that of Lemma 3.7. Recall that $m_n: x\mapsto\mathbb{E}\left[r_n(Z)\mid X=x\right]$. Starting as in the proof of the latter lemma, we can write

$$\begin{aligned}
&\max_{1\leq i\leq n}\left\|\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K\left(X_i,X_j,b_n\right)r_n(Z_j)-m_n(X_i)f_X(X_i)\right\|\\
\leq&\max_{1\leq i\leq n}\left\|\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K\left(X_i,X_j,b_n\right)r_n(Z_j)-\mathbb{E}\left[\frac{1}{b_n^{d_x}}K\left(X_i,X,b_n\right)r_n(Z)\mid X_i\right]\right\|\\
&+\sup_{x\in[0,1]^{d_x}}\left\|\mathbb{E}\left[\frac{1}{b_n^{d_x}}K\left(x,X,b_n\right)r_n(Z)\right]-m_n(x)f_X(x)\right\|\\
=:&A_1+A_2.
\end{aligned}$$

We first observe that A_1 can be further bounded by

$$\begin{aligned}
&\frac{\overline{K}^{d_x}}{nb_n^{d_x}}\left(\max_{1\leq i\leq n}\|r_n(Z_i)\|+\mathbb{E}\left[\|r_n(Z)\|\right]\right)\\
&+\max_{1\leq i\leq n}\left\|\frac{1}{(n-1)b_n^{d_x}}\sum_{j\neq i}K\left(X_i,X_j,b_n\right)r_n(Z_j)-\mathbb{E}\left[\frac{1}{b_n^{d_x}}K\left(X_i,X,b_n\right)r_n(Z)\mid X_i\right]\right\|\\
=:&A_3+A_4.
\end{aligned}$$

Given the assumptions of the lemma, $A_3=O_P(1/nb_n^{d_x})$ by Markov's inequality. When $r_n(z)=\rho(z,\Pi_n h_0)$, we can apply Lemma 3.10 with $f_n(z)=\rho_l(z,\Pi_n h_0)$ for $l\in\{1,...,d\}$ to control A_4. When $r_n(z)=\rho(z,\Pi_n h_0)\rho(z,\Pi_n h_0)^t$, we can also use Lemma 3.10 with $f_n(z)=\rho_l(z,\Pi_n h_0)\rho_{l'}(z,\Pi_n h_0)$ for every $(l,l')\in\{1,...,d\}^2$ to control A_4. In both cases, we obtain $A_4=O_P\left(\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}}\right)$. The control of A_2 is similar to the control of A_2 in the proof of Lemma 3.7 and is thus omitted. We can claim that $A_2=O(b_n^s)$.

Gathering all the intermediary results, we conclude

$$\max_{1\leq i\leq n}\left\|\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K\left(X_i,X_j,b_n\right)r_n(Z_j)-m_n(X_i)f_X(X_i)\right\|=O_P\left(\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}}+b_n^s\right).$$

3.6.2.8 Proof of Lemma 3.9

Let $\overline{f}_{Z_h}:=\sup_{z_h\in\mathcal{Z}_h}f_{Z_h}(z_h)$. First we observe that for every $M_0>0$, $\mathcal{H}^{M_0}$ is a bounded subset of $\mathcal{H}$: since $\overline{f}_{Z_h}<+\infty$, we can claim that for every $h\in\mathcal{H}^{M_0}$, we have $\|h\|_2^2\leq\max\left\{\overline{f}_{Z_h},1\right\}\|h\|_{L_2(leb)}\leq\max\left\{\overline{f}_{Z_h},1\right\}M_0$. This, $m-d_{z_h}/2>0$ and $\|\langle Z_h\rangle^\gamma\|_2<+\infty$ for some $\gamma>0$ ensure that Corollary 4 in [114] is applicable (with $\beta=0$). This corollary states that for every $M_0>0$ the following holds

$$N_{[\,]}\left(\epsilon,\mathcal{H}^{M_0},L_2(P)\right)<+\infty\text{ for every }\epsilon>0.$$

Using the fact that bracketing numbers are larger than covering numbers, we obtain

$$N\left(\epsilon, \mathcal{H}^{M_0}, L_2(P)\right) < +\infty \text{ for every } \epsilon > 0,$$

i.e $\mathcal{H}^{M_0}$ is a totally bounded subset of $L_2(P)$ for every $M_0 > 0$. The closure of a totally bounded set is itself totally bounded,[6] which is enough to claim that for every $M_0 > 0$, $\overline{\mathcal{H}}^{M_0}$ is a close and totally bounded subset of $L_2(P)$, that is to say a compact subset of $L_2(P)$.

3.6.2.9 Proof of Lemma 3.10

To avoid notational burden, we give the result in the simplified case where $K(x, y, b_n) = \prod_{t=1}^{d_x} \widetilde{K}\left(\frac{x^{(t)}-y^{(t)}}{b_n}\right) =: K\left(\frac{x-y}{b_n}\right)$. To handle the actual $K(x, y, b_n)$ function we consider, the steps are the same.

Let $\widetilde{f} : z \mapsto 1$ and $\mathcal{F}_n := \{f_n\} \cup \{\widetilde{f}\}$. We remark that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{x\in[0,1]^{d_x}}\left|\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n} K\left(\frac{x-X_j}{b_n}\right)h_n(Z_j) - \mathbb{E}\left(\frac{1}{b_n^{d_x}}K\left(\frac{x-X}{b_n}\right)h_n(Z)\right)\right|\right] \\
\leq &\frac{1}{b_n^{d_x}}\mathbb{E}\left[\sup_{(x,f)\in[0,1]^{d_x}\times\mathcal{F}_n}\left|\frac{1}{n}\sum_{j=1}^{n} K\left(\frac{x-X_j}{b_n}\right)f(Z_j) - \mathbb{E}\left(K\left(\frac{x-X}{b_n}\right)f(Z)\right)\right|\right]. \qquad (3.48)
\end{aligned}$$

The class of functions $\mathcal{G}_n := \left\{K\left(\frac{x-\cdot}{b_n}\right)f(\cdot) : (x, f) \in [0,1]^{d_x} \times \mathcal{F}_n\right\}$ admits an envelope $G_n : z \mapsto \overline{K}^{d_x}\max\{f_n(z), 1\}$ which verifies $\sup_{n\geq 1}\mathbb{E}\left[G_n(Z)^2\right] < +\infty$ by assumption. The class $\mathcal{G}_n$ can also be viewed as the product between $\mathcal{K}_n := \left\{K\left(\frac{x-\cdot}{b_n}\right) : x \in [0,1]^{d_x}\right\}$ (with envelope $K_n : x \mapsto \overline{K}^{d_x}$) and $\mathcal{F}_n$ (with envelope $F_n : z \mapsto \max\{f_n(z), 1\}$). Corollary 7(i) in [98] ensures that for every $\epsilon > 0$

$$\begin{aligned}
&\sup_Q N\left(2\epsilon\left\|G_n\right\|_{L_2(Q)}, \mathcal{G}_n, L_2(Q)\right) \\
\leq &\sup_Q N\left(\epsilon\left\|K_n\right\|_{L_2(Q)}, \mathcal{K}_n, L_2(Q)\right)\sup_Q N\left(\epsilon\left\|F_n\right\|_{L_2(Q)}, \mathcal{F}_n, L_2(Q)\right),
\end{aligned}$$

where the supremum is taken over all discrete probability measures.

Let $\widetilde{\mathcal{K}}_n := \left\{\widetilde{K}\left(\frac{x-\cdot}{b_n}\right) : x \in [0,1]\right\}$. Assumption 3.4(v) and the definition of a VC-type class of functions imply there exist positive constants A and v independent of n such that for every $\epsilon > 0$

$$\sup_Q N\left(\epsilon\overline{K}, \widetilde{\mathcal{K}}_n, L_2(Q)\right) \leq \left(\frac{A}{\epsilon}\right)^v.$$

Applying Proposition 5 in [98] with $\phi : x \in \mathbb{R}^{d_x} \mapsto \prod_{t=1}^{d_x} x^{(t)}$, $k = d_x$ and $(\mathcal{F}_j)_{j=1}^{d_x}$ replaced with $\left(\left\{\widetilde{K}\left(\frac{x^{(t)}-\cdot}{b_n}\right) : x^{(t)} \in [0,1]\right\}\right)_{t=1}^{d_x}$, we can write for every $\epsilon > 0$

$$\sup_Q N\left(\epsilon\left\|K_n\right\|_{L_2(Q)}, \mathcal{K}_n, L_2(Q)\right) \leq \left(\frac{d_x A}{\overline{K}^{d_x}\epsilon}\right)^{d_x v}.$$

We let $A_1 = \frac{d_x A}{\overline{K}^{d_x}}$ and $v_1 = d_x v$. Using the last inequality and the fact that the cardinal of $\mathcal{F}_n$ is 2, we get for every $\epsilon > 0$

$$\sup_Q N\left(\epsilon\left\|G_n\right\|_{L_2(Q)}, \mathcal{G}_n, L_2(Q)\right) \leq 2\left(\frac{2A_1}{\epsilon}\right)^{v_1},$$

[6] It is not difficult to see that for every positive M_0 and ϵ: $N\left(\epsilon, \overline{\mathcal{H}}^{M_0}, L_2(P)\right) \leq N\left(\epsilon/2, \mathcal{H}^{M_0}, L_2(P)\right)$.

which combined with Corollary 5.1 in [43] enables us to write

$$\begin{aligned}
&\mathbb{E}\left[\sup_{(x,f)\in[0,1]^{d_x}\times\mathcal{F}_n}\left|\frac{1}{n}\sum_{j=1}^{n}K\left(\frac{x-X_j}{b_n}\right)f(Z_j)-\mathbb{E}\left(K\left(\frac{x-X}{b_n}\right)f(Z)\right)\right|\right]\\
\leq&\frac{C_1}{\sqrt{n}}\left\{\sqrt{vB_n\log\left(\frac{2^{1/v_1}2A_1\sup_{n\geq1}\mathbb{E}\left[G_n(Z)^2\right]}{\sqrt{B_n}}\right)}\right.\\
&\left.+\frac{v_1\sup_{n\geq1}\mathbb{E}\left[\max_{1\leq j\leq n}G_n(Z_j)^2\right]}{\sqrt{n}}\log\left(\frac{2^{1/v_1}2A_1\sup_{n\geq1}\mathbb{E}\left[G_n(Z)^2\right]}{\sqrt{B_n}}\right)\right\},
\end{aligned}\tag{3.49}$$

where B_n is any number between $\sup_{(x,f)\in[0,1]^{d_x}\times\mathcal{F}_n}\mathbb{E}\left[K\left(\frac{x-X}{b_n}\right)^2f(Z)^2\right]$ and $\mathbb{E}\left[G_n(X,Z)^2\right]$.

We remark that

$$\begin{aligned}
&\sup_{(x,f)\in[0,1]^{d_x}\times\mathcal{F}_n}\mathbb{E}\left[K\left(\frac{x-X}{b_n}\right)^2f(Z)^2\right]\\
\leq&b_n^{d_x}\sup_{n\geq1}\sup_{x\in[0,1]^{d_x}}\mathbb{E}\left[\max\{f_n(Z)^2,1\}\mid X=x\right]\overline{f}_X\int_{[-1,1]^{d_x}}K(u)^2du=b_n^{d_x}C_2,
\end{aligned}$$

which is smaller than $\mathbb{E}\left[G_n(X,Z)^2\right]$ for n large enough since $b_n^{d_x}\to0$ as n goes to $+\infty$ while $\inf_{n\geq1}\mathbb{E}\left[G_n(X,Z)^2\right]>0$ by construction. We can therefore pick $B_n=b_n^{d_x}C_2$ in (3.49).

Combining (3.48) and (3.49)

$$\begin{aligned}
&\mathbb{E}\left[\sup_{x\in[0,1]^{d_x}}\left|\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K\left(\frac{x-X_j}{b_n}\right)f_n(Z_j)-\mathbb{E}\left(\frac{1}{b_n^{d_x}}K\left(\frac{x-X}{b_n}\right)f_n(Z)\right)\right|\right]\\
\leq&\frac{C_1}{\sqrt{n}b_n^{d_x}}\left\{\sqrt{v_1C_2b_n^{d_x}\log\left(C_3b_n^{-d_x/2}\right)}+\frac{C_4}{\sqrt{n}}\log\left(C_3b_n^{-d_x/2}\right)\right\},
\end{aligned}$$

where $C_3=\frac{2^{1/v_1}\sqrt{2}A_1\sup_{n\geq1}\mathbb{E}\left[G_n(Z_1)^2\right]}{\sqrt{C_2}}$ and $C_4=v_1\sup_{n\geq1}\mathbb{E}\left[\max_{1\leq j\leq n}G_n(Z_j)^2\right]$. We can simplify the upper bound even further using the fact that $\frac{1}{n-1}\leq\frac{2}{n}$ and for n large enough $\log\left(C_3b_n^{-d_x/2}\right)\leq d_x\left|\log b_n\right|$ and $\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}}\geq\frac{|\log b_n|}{nb_n^{d_x}}$

$$\begin{aligned}
&\mathbb{E}\left[\sup_{x\in[0,1]^{d_x}}\left|\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K\left(\frac{x-X_j}{b_n}\right)f_n(Z_j)-\mathbb{E}\left(\frac{1}{b_n^{d_x}}K\left(\frac{x-X}{b_n}\right)f_n(Z)\right)\right|\right]\\
\leq&C_5\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}},
\end{aligned}$$

with $C_5=2C_1\max\left\{\sqrt{d_xv_1C_2},d_xC_4\right\}$. This is enough to conclude.

To see that the result is still true when $\mathcal{K}_n$ refers to $\left\{\left|K\left(\frac{x-\cdot}{b_n}\right)\right|:x\in[0,1]^{d_x}\right\}$, we observe that $\overline{K}^{d_x}$ remains a valid envelope and for every $(x_1,x_2)\in[0,1]^{d_x}\times[0,1]^{d_x}$ and every probability measure Q on $[0,1]^{d_x}$ (endowed with its Borel sigma-algebra)

$$\begin{aligned}
&\int_{[0,1]^{d_x}}\left|\left|K\left(\frac{x_1-u}{b_n}\right)\right|-\left|K\left(\frac{x_2-u}{b_n}\right)\right|\right|^2dQ(u)\\
\leq&\int_{[0,1]^{d_x}}\left|K\left(\frac{x_1-u}{b_n}\right)-K\left(\frac{x_2-u}{b_n}\right)\right|^2dQ(u).
\end{aligned}$$

3.6.2.10 Proof of Lemma 3.11

Let $(X,Z) \sim P$ and $(X,Z) \perp\!\!\!\perp (X_i,Z_i)_{i=1}^n$ and $f(z) = \sup_{h\in\mathcal{G}} \|\rho(z,h)\|^q$. We first note

$$\begin{aligned}
&\mathbb{E}\left[\max_{1\le i\le n} \frac{1}{nb_n^{d_x}} \sum_{j=1}^n |K_{ij}|\, f(Z_j)\right] \\
\le&\mathbb{E}\left[\max_{1\le i\le n} \left|\frac{1}{nb_n^{d_x}} \sum_{j=1}^n |K_{ij}|\, f(Z_j) - \mathbb{E}\left[\frac{1}{b_n^{d_x}} |K(X_i,X,b_n)|\, f(Z_j) \mid X_i\right]\right|\right] \\
&+ \sup_{x\in[0,1]^{d_x}} E\left[\frac{1}{b_n^{d_x}} |K(x,X,b_n)|\, f(Z_j)\right] \\
\le&\mathbb{E}\left[\sup_{x\in[0,1]^{d_x}} \left|\frac{1}{(n-1)b_n^{d_x}} \sum_{j=2}^n |K(x,X_j,b_n)|\, f(Z_j) - \mathbb{E}\left[\frac{1}{b_n^{d_x}} |K(x,X,b_n)|\, f(Z_j)\right]\right|\right] \\
&+ \frac{C_1}{nb_n^{d_x}} + \sup_{x\in[0,1]^{d_x}} E\left[\frac{1}{b_n^{d_x}} |K(x,X,b_n)|\, f(Z_j)\right],
\end{aligned}$$

with $C_1 := 2\overline{K}^{d_x}\mathbb{E}[f(Z)]$.

Using Lemma 3.10 with $f_n(z) = f(z)$, we can claim that there exists $N_1 \ge 1$ such that for every $n > N_1$

$$\begin{aligned}
&\mathbb{E}\left[\sup_{x\in[0,1]^{d_x}} \left|\frac{1}{(n-1)b_n^{d_x}} \sum_{j=2}^n |K(x,X_j,b_n)|\, f(Z_j) - \mathbb{E}\left(\frac{1}{b_n^{d_x}} |K(x,X,b_n)|\, f(Z)\right)\right|\right] \\
\le& C_2\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}},
\end{aligned}$$

for some constant C_2 that depends on d_x, $K(\cdot)$, P and $f(\cdot)$.

By a change of variable and Assumption 3.4(iii)-(iv), we also get $\sup_{x\in[0,1]^{d_x}} \mathbb{E}\left(\frac{1}{b_n^{d_x}} |K(x,X,b_n)|\, f(Z)\right) \le C_3$. Those two results and $\max\left\{\frac{|\log b_n|}{nb_n^{d_x}}, \frac{1}{nb_n^{d_x}}\right\} = o(1)$ imply that there exists $N \ge N_1$ such that for every $n > N$

$$\mathbb{E}\left[\max_{1\le i\le n} \frac{1}{nb_n^{d_x}} \sum_{j=1}^n |K_{ij}|\, f(Z_j)\right] \le \frac{C_1}{nb_n^{d_x}} + C_2\sqrt{\frac{|\log b_n|}{nb_n^{d_x}}} + C_3 \le 3\max\{C_1,C_2,C_3\}.$$

3.6.2.11 Proof of Lemma 3.12

We only give the proof in the case $r(z,h) = \rho(z,h)$. The proof for $r(z,h) = \rho(z,h)\rho(z,h)^t$ is exactly the same, up to notational changes.

By the triangle inequality, a convexity argument and Assumption 3.1

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\sup_{h\in\mathcal{H}^{M_0}}\left\|\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K\left(X_i,X_j,b_n\right)\rho\left(Z_j,h\right)-\mathbb{E}\left(\frac{1}{b_n^{d_x}}K\left(X_i,X,b_n\right)\rho\left(Z,h\right)\mid X_i\right)\right\|^2\right]\\
\leq& d\max_{1\leq l\leq d}\mathbb{E}\left[\sup_{h\in\mathcal{H}^{M_0}}\left|\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K\left(X_1,X_j,b_n\right)\rho_l(Z_j,h)-\mathbb{E}\left(\frac{1}{b_n^{d_x}}K\left(X_1,X,b_n\right)\rho_l(Z,h)\mid X_1\right)\right|^2\right]\\
\leq&\frac{2d}{n^2b_n^{2d_x}}\max_{1\leq l\leq d}\mathbb{E}\left[\sup_{h\in\mathcal{H}^{M_0}}\left|K\left(0\right)\rho_l(Z_1,h)-\mathbb{E}\left(K\left(X_1,X,b_n\right)\rho_l(Z,h)\mid X_1\right)\right|^2\right]\\
&+\frac{2d}{n^2b_n^{2d_x}}\max_{1\leq l\leq d}\mathbb{E}\left[\sup_{h\in\mathcal{H}^{M_0}}\left|\sum_{j=2}^{n}K\left(X_1,X_j,b_n\right)\rho_l(Z_j,h)-\mathbb{E}\left(K\left(X_1,X,b_n\right)\rho_l(Z,h)\mid X_1\right)\right|^2\right]\\
=:&A_1+A_2.
\end{aligned}\tag{3.50}$$

Control of A_1

Under Assumptions 3.1, 3.2 and 3.4(i)-(iii) and a convexity argument

$$A_1\leq\frac{8d}{n^2b_n^{2d_x}}\overline{K}^{d_x}\mathbb{E}\left[\sup_{h\in\mathcal{H}^{M_0}}\left\|\rho(Z,h)\right\|^2\right]=\frac{C_1d}{n^2b_n^{2d_x}}\tag{3.51}$$

with $C_1=8\overline{K}^{d_x}\mathbb{E}\left[\sup_{h\in\mathcal{H}^{M_0}}\left\|\rho(Z,h)\right\|^2\right]$.

Control of A_2

The term A_2 is upper bounded by

$$\begin{aligned}
&\frac{2d(n-1)^2}{n^2b_n^{2d_x}}\\
&\times\max_{1\leq l\leq d}\mathbb{E}\left[\sup_{h\in\mathcal{H}^{M_0}}\left|\frac{1}{(n-1)}\sum_{j=2}^{n}K\left(X_1,X_j,b_n\right)\rho_l(Z_j,h)-\mathbb{E}\left(K\left(X_1,X,b_n\right)\rho_l(Z,h)\mid X_1\right)\right|^2\right]\\
\leq&\frac{2d}{b_n^{2d_x}}\max_{1\leq l\leq d}\mathbb{E}\left[\sup_{h\in\mathcal{H}^{M_0}}\left|\frac{1}{(n-1)}\sum_{j=2}^{n}K\left(X_1,X_j,b_n\right)\rho_l(Z_j,h)\right.\\
&\left.-\mathbb{E}\left(K\left(X_1,X,b_n\right)\rho_l(Z,h)\mid X_1\right)\right|^2\right].
\end{aligned}\tag{3.52}$$

Note that under the conditional distribution $P^{(X_j,Z_j)_{j=2}^n|X_1=x_1}$

$$\frac{1}{n-1}\sum_{j=2}^{n}K\left(x_1,X_j,b_n\right)\rho_l(Z_j,h)-\mathbb{E}\left(K\left(x_1,X,b_n\right)\rho_l(Z,h)\mid X_1=x_1\right)$$

is a sum of $(n-1)$ centered and *i.i.d* random variables, indexed by the class of functions $\mathcal{F}_{n,x_1}^l:=\left\{(x,z)\mapsto K\left(x_1,x,b_n\right)\rho_l(z,h):h\in\mathcal{H}^{M_0}\right\}$. This class of functions depends on n, is parameterized by x_1 and M_0, and has an envelope $F_{n,x_1}^l(x,z)=$
$|K\left(x_1,x,b_n\right)|\left(\sup_{h\in\mathcal{H}^{M_0}}||\rho(z,h)||+1\right)$ with finite $L_2\left(P^{(Z,X)|X_1=x_1}\right)$ norm (here (Z,X) stands for an *i.i.d* copy of (Z_1,X_1)).

Let $P^{\cdot|X_1}$ stand for $P^{(Z,X)|X_1}$. Observe that by a change of variable and Assumption 3.4(iii)-(iv)

$$\begin{aligned}
&\mathbb{E}\left[F_{n,x_1}^l(X,Z)^2\mid X_1=x_1\right]\\
\leq&\overline{f}_X\sup_{x\in[0,1]^{d_x}}\mathbb{E}\left[\sup_{h\in\mathcal{H}^{M_0}}\left(||\rho(Z,h)||+1\right)^2\mid X=x\right]\int_{[0,1]^{d_x}}K\left(x_1,x,b_n\right)^2dx\\
\leq&b_n^{d_x}\overline{f}_X\sup_{x\in[0,1]^{d_x}}\mathbb{E}\left[\sup_{h\in\mathcal{H}^{M_0}}\left(||\rho(Z,h)||+1\right)^2\mid X=x\right]\widetilde{C},
\end{aligned}$$

where $\widetilde{C}$ depends on $K(\cdot)$.

The upper bound is finite under Assumptions 3.2. This implies that for some $C_2 > 0$

$$\sup_{x_1 \in [0,1]^{d_x}} \left|\left|F^l_{n,x_1}\right|\right|_{L_2\left(P^{\cdot|X_1=x_1}\right)} \leq C_2 b_n^{d_x/2}. \tag{3.53}$$

As a result, we can apply Remark 3.5.14 and Equation (3.214) that follows in [79] plus Theorem 3.1.22 from the same book to upper bound the expectation of

$$\sup_{h \in \mathcal{H}^{M_0}} \left| \frac{1}{n-1} \sum_{j=2}^{n} K\left(X_1, X_j, b_n\right) \rho_l(Z_j, h) - \mathbb{E}\left(K\left(X_1, X, b_n\right) \rho_l(Z, h) \mid X_1\right) \right|$$

conditionally on X_1. We obtain (using also that $P^{(X,Z)|X_1=x_1} = P^{(X,Z)}$)

$$\begin{aligned}
&\mathbb{E}\left[\sup_{h \in \mathcal{H}^{M_0}} \left| \frac{1}{n-1} \sum_{j=2}^{n} K\left(X_1, X_j, b_n\right) \rho_l(Z_j, h) - \mathbb{E}\left(K\left(X_1, X, b_n\right) \rho_l(Z, h) \mid X_1\right) \right|^2 \mid X_1 \right] \\
\leq & \frac{C_3}{n-1} \left\{ \frac{1}{n-1} \mathbb{E}\left[\max_{j \in \{2,\ldots,n\}} K\left(X_1, X_j, b_n\right)^2 \sup_{h \in \mathcal{H}^{M_0}} \left\|\rho(Z_j, h)\right\|^2 \mid X_1 \right] \right. \\
& \left. + \frac{1}{n-1} \mathbb{E}\left[\sup_{h \in \mathcal{H}^{M_0}} \left| \sum_{j=2}^{n} K\left(X_1, X_j, b_n\right) \rho_l(Z_j, h) - \mathbb{E}\left(K\left(X_1, X, b_n\right) \rho_l(Z, h) \mid X_1\right) \right| \mid X_1 \right]^2 \right\} \\
\leq & \frac{C_4}{n-1} \left\{ A_n^2 + \mathbb{E}\left[K\left(X_1, X, b_n\right)^2 \sup_{h \in \mathcal{H}^{M_0}} \left\|\rho(Z, h)\right\|^2 \mid X_1 \right] \right\} \\
\leq & \frac{C_4}{n-1} \left\{ A_n^2 + \sup_{x_1 \in [0,1]^{d_x}} \left|\left|F^l_{n,x_1}\right|\right|^2_{L_2(P^{X,Z})} \right\},
\end{aligned}$$

where C_3 and C_4 are universal constants, $A_n = \left|\left|F^l_{n,X_1}(X,Z)\right|\right|_{L_2\left(P^{\cdot|X_1}\right)} J_l(n, X_1, M_0, P)$[7] and

$$J_l(n, x_1, M_0, P) := \int_0^1 \sqrt{1 + \log N_{[\,]}\left(\epsilon \left|\left|F^l_{n,x_1}(X,Z)\right|\right|_{L_2(P^{X,Z})}, \mathcal{F}^l_{n,x_1}, L_2\left(P^{X,Z}\right)\right)} d\epsilon.$$

The upper bound is valid P^{X_1}-*a.s* and we can integrate on each side of the inequality with respect to P^{X_1} to obtain

$$\begin{aligned}
&\mathbb{E}\left[\sup_{h \in \mathcal{H}^{M_0}} \left| \frac{1}{n-1} \sum_{j=2}^{n} K\left(X_1, X_j, b_n\right) \rho_l(Z_j, h) - \mathbb{E}\left(K\left(X_1, X, b_n\right) \rho_l(Z, h) \mid X_1\right) \right|^2 \right] \\
\leq & \frac{C_4}{n-1} \left\{ \mathbb{E}\left[A_n^2\right] + \sup_{x_1 \in [0,1]^{d_x}} \left|\left|F^l_{n,x_1}\right|\right|^2_{L_2\left(P^{\cdot|X_1=x_1}\right)} \right\}.
\end{aligned} \tag{3.54}$$

Assumption 3.6 and (3.53) further ensure that for some $C_5 > 0$

$$\sup_{x_1 \in [0,1]^{d_x}} \left\{ \left|\left|F^l_{n,x_1}(X_1, X, Z)\right|\right|_{L_2\left(P^{\cdot|X_1=x_1}\right)} J_l(n, x_1, M_0, P) \right\} \leq C_5 b_n^{d_x/2}. \tag{3.55}$$

Combine (3.52), (3.53), (3.54), (3.55) and $n/(n-1) \leq 2$ whenever $n \geq 2$ to claim that for every $n \geq 2$

$$A_2 \leq \frac{dC_6}{n b_n^{d_x}}, \tag{3.56}$$

with $C_6 = 4C_4(C_2^2 + C_5^2)$.

[7] Note that if the random quantity $J(n, X_1, M_0, P)$ is not Borel-measurable, then $\mathbb{E}\left[A_n^2\right]$ has to be replaced with an outer expectation.

Conclusion

Let $N \geq 1$ be such that $\frac{1}{nb_n^{d_x}} \leq 1$ for every $n > N$. This N exists since $nb_n^{d_x} \to +\infty$. Combine (3.50), (3.51) and (3.56) to conclude that for every $n > N$

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\sup_{h\in\mathcal{H}^{M_0}}\left\|\frac{1}{nb_n^{d_x}}\sum_{j=1}^{n}K\left(X_i,X_j,b_n\right)\rho\left(Z_j,h\right)-\mathbb{E}\left(\frac{1}{b_n^{d_x}}K\left(X_i,X,b_n\right)\rho\left(Z,h\right)\mid X_i\right)\right\|^2\right] \leq \frac{2d\max\{C_1,C_6\}}{nb_n^{d_x}}.$$

3.6.2.12 Proof of Lemma 3.13

Show that $\sup_{x_1\in[0,1]^{d_x}} J_l(n,x_1,M_0,P) < +\infty$ in the Lipschitz case

We focus on the case: $|\rho_l(z,h_1)-\rho_l(z,h_2)| \leq L(Z)\,|h_1(Z_h)-h_2(Z_h)|$ for every $M_0 > 0$, $(h_1,h_2) \in \mathcal{H}^{M_0}\times\mathcal{H}^{M_0}$ and $l \in \{1,...,d\}$.

For every pair $(h_1,h_2) \in \mathcal{H}^{M_0}\times\mathcal{H}^{M_0}$

$$\begin{aligned}&|K\left(x_1,x,b_n\right)\times\rho_l(z,h_1)-K\left(x_1,x,b_n\right)\times\rho_l(z,h_2)|\\ \leq&|K\left(x_1,x,b_n\right)|\times L(z)\,|h_1(z_h)-h_2(z_h)| \leq |K\left(x_1,x,b_n\right)|\times L(z)\langle z_h\rangle^{\gamma}\,||h_1-h_2||_{\infty,\gamma}\end{aligned}$$

where $||h||_{\infty,\gamma} := \sup_{z_h\in\mathcal{Z}_h}|h(z_h)\times\langle z_h\rangle^{-\gamma}|$ acts as a weighted sup-norm on $\mathcal{H}$ and $\langle z_h\rangle = (1+||z_h||^2)^{1/2}$.

Following steps in the proofs of Theorem 1 and Corollary 4 in [114] and in particular Equation (3) on page 184, for every positive M_0 and ϵ and every γ such that $\gamma > m - \frac{d_{z_h}}{2} > 0$, we can build a finite number of ϵ balls under $||.||_{\infty,\gamma}$ to cover $\mathcal{H}^{M_0}$ such that

$$N(\epsilon) := N\left(\epsilon,\mathcal{H}^{M_0},||.||_{\infty,\gamma}\right) \leq K(M_0)\exp\left\{\epsilon^{-d_{z_h}/m}\right\},$$

where $K(M_0)$ is some positive constant that depends on M_0.

Denote $\{h_i\}_{i=1}^{N(\epsilon)}$ the centers of each ϵ ball. By construction of those balls, for every $\epsilon > 0$ and $h \in \mathcal{H}^{M_0}$, there exists h_i such that $||h-h_i||_{\infty,\gamma} \leq \epsilon$. This implies that for every $\epsilon > 0$ and $h \in \mathcal{H}^{M_0}$, there exists h_i such that uniformly in $x \in [0,1]^{d_x}$

$$\begin{aligned}&K\left(x_1,x,b_n\right)\times\rho_l(z,h_i)-\epsilon\,|K\left(x_1,x,b_n\right)|\times L(z)\langle z_h\rangle^{\gamma}\\ \leq&K\left(x_1,x,b_n\right)\times\rho_l(z,h) \leq K\left(x_1,x,b_n\right)\times\rho_l(z,h_i)+\epsilon\,|K\left(x_1,x,b_n\right)|\times L(z)\langle z_h\rangle^{\gamma}.\end{aligned} \tag{3.57}$$

The relation in (3.57) shows that

$$\begin{aligned}&\Big\{K\left(x_1,x,b_n\right)\times\rho_l(z,h_i)-\epsilon\,|K\left(x_1,x,b_n\right)|\times L(z)\langle z_h\rangle^{\gamma},\\ &K\left(x_1,x,b_n\right)\times\rho_l(z,h_i)+\epsilon\,|K\left(x_1,x,b_n\right)|\times L(z)\langle z_h\rangle^{\gamma}\Big\}_{i=1}^{N(\epsilon)}\end{aligned}$$

is a valid set of brackets of $\mathcal{F}_{n,x_1}^l$ with $L_2\left(P^{X,Z}\right)$ size $2\epsilon\sqrt{\mathbb{E}\left[\left(K\left(x_1,X,b_n\right)L(Z)\langle Z_h\rangle^{\gamma}\right)^2\right]}$. This implies

$$N_{[\,]}\left(2\epsilon\sqrt{\mathbb{E}\left[\left(K\left(x_1,X,b_n\right)L(Z)\langle Z_h\rangle^{\gamma}\right)^2\right]},\mathcal{F}_{n,x_1}^l,L_2\left(P^{X,Z}\right)\right) \leq N\left(\epsilon,\mathcal{H}^{M_0},||.||_{\infty,\gamma}\right),$$

and

$$N_{[\,]}\left(\epsilon\left|\left|F_{n,x_1}^l\right|\right|_{L_2(P^{X,Z})},\mathcal{F}_{n,x_1}^l,L_2\left(P^{X,Z}\right)\right) \leq N\left(\epsilon K(x_1,M_0,\gamma),\mathcal{H}^{M_0},||.||_{\infty,\gamma}\right),$$

where $K(x_1, M_0, \gamma) = \frac{||F^l_{n,x_1}||_{L_2(P^{X,Z})}}{2\sqrt{\mathbb{E}\left[\left(K(x_1,X,b_n)L(Z)\langle Z_h\rangle^\gamma\right)^2\right]}}$.

Since F^l_{n,x_1} is essentially bounded from below by $|K(x_1, x, b_n)|$ and Assumption 3.7(i) entails $\sup_{x\in[0,1]^{d_x}} \mathbb{E}\left[\left(L(Z)\langle Z_h\rangle^\gamma\right)^2 \mid X = x\right] < +\infty$, we arrive at

$$\begin{aligned} K(x_1, M_0, \gamma) \geq & 0.5\sqrt{\frac{\mathbb{E}\left[|K(x_1, X, b_n)|^2\right]}{\mathbb{E}\left[|K(x_1, X, b_n)|^2 \mathbb{E}\left[\left(L(Z)\langle Z_h\rangle^\gamma\right)^2 \mid X\right]\right]}} \\ \geq & 0.5\left(\sup_{x\in[0,1]^{d_x}} \mathbb{E}\left[\left(L(Z)\langle Z_h\rangle^\gamma\right)^2 \mid X = x\right]\right)^{-1/2} = K(\gamma) \end{aligned}$$

which does not depend on x_1 anymore.

We conclude that uniformly in x_1

$$\sqrt{\log N_{[\,]}\left(\epsilon\left|\left|F^l_{n,x_1}\right|\right|_{L_2(P^{X,Z})}, \mathcal{F}^l_{n,x_1}, L_2\left(P^{X,Z}\right)\right)} \leq K(M_0)(\epsilon K(\gamma))^{-d_{z_h}/2m}.$$

We can thus see that in the Lipschitz case, whenever $m/d_{z_h} > 1/2$,

$$\sup_{x_1\in[0,1]^{d_x}} J_l(n, x_1, M_0, P) < +\infty.$$

Show that $\sup_{x_1\in[0,1]^{d_x}} J_{l,l'}(n, x_1, M_0, P) < +\infty$ in the Lipschitz case

We follow the same lines as those that enabled us to conclude $\sup_{x_1\in[0,1]^{d_x}} J_l(n, x_1, M_0, P) < +\infty$. For every $M_0 > 0$, $(l, l') \in \{1, ..., d\}^2$, $x_1 \in [0,1]^{d_x}$ and $(h_1, h_2) \in \mathcal{H}^{M_0} \times \mathcal{H}^{M_0}$

$$\begin{aligned} & |K(x_1, x, b_n) \times \rho_l(z, h_1)\rho_{l'}(z, h_1) - K(x_1, x, b_n) \times \rho_l(z, h_2)\rho_{l'}(z, h_2)| \\ \leq & 2|K(x_1, x, b_n)| \times \sup_{h\in\mathcal{H}^{M_0}} \|\rho(z, h)\|L(z)\langle z_h\rangle^\gamma \, ||h_1 - h_2||_{\infty,\gamma}. \end{aligned}$$

For every $\epsilon > 0$, let $\{h_i\}_{i=1}^{N(\epsilon)}$ be as defined in the previous subsection. We infer from the last inequality that

$$\begin{aligned} \Big\{ & K(x_1, x, b_n) \times \rho_l(z, h_i)\rho_{l'}(z, h_i) - \epsilon 2|K(x_1, x, b_n)| \times \sup_{h\in\mathcal{H}^{M_0}} \|\rho(z, h)\|L(z)\langle z_h\rangle^\gamma, \\ & K(x_1, x, b_n) \times \rho_l(z, h_i)\rho_{l'}(z, h_i) + \epsilon 2|K(x_1, x, b_n)| \times \sup_{h\in\mathcal{H}^{M_0}} \|\rho(z, h)\|L(z)\langle z_h\rangle^\gamma \Big\}_{i=1}^{N(\epsilon)} \end{aligned}$$

is a valid set of brackets of $\mathcal{F}^{l,l'}_{n,x_1}$ with $L_2\left(P^{X,Z}\right)$ size

$$4\epsilon\sqrt{\mathbb{E}\left[\left(K(x_1, X, b_n) \sup_{h\in\mathcal{H}^{M_0}} \|\rho(Z, h)\|L(Z)\langle Z_h\rangle^\gamma\right)^2\right]}.$$

Still following the steps in the last subsection, we conclude that as long as $m/d_{z_h} > 1/2$,

$$\sup_{x\in[0,1]^{d_x}} \mathbb{E}\left[\sup_{h\in\mathcal{H}^{M_0}} \|\rho(Z, h)\|^4 \mid X = x\right] < +\infty \ \forall M_0 > 0$$

and $\sup_{x\in[0,1]^{d_x}} \mathbb{E}\left[\left(L(Z)\langle Z_h\rangle^\gamma\right)^4 \mid X = x\right] < +\infty$, the result holds.

Show that $\sup_{x_1\in[0,1]^{d_x}} J_l(n, x_1, M_0, P) < +\infty$ in the NPQIV case

We now let $\rho(z, h) = \mathbb{1}\{z_o \leq h(z_h)\} - \tau$ for some $\tau \in]0, 1[$. The method of proof is borrowed from Babii & Florens (2017). Note that here there is a single moment condition so that we can drop the dependence of $J_l(n, x_1, M_0, P)$ and $\mathcal{F}^l_{n,x_1}$ on l.

We build a minimal ϵ covering of $\mathcal{H}^{M_0}$ under the $||.||_{\infty,\gamma}$ norm and denote $\{h_i\}_{i=1}^{N(\epsilon)}$ the family of centers of balls. As explained in [114], for every $h \in \mathcal{H}^{M_0}$, there exists h_i such that for every $z_h \in \mathcal{Z}_h$, $h_i(z_h) - \epsilon\langle z_h\rangle^\gamma \leq h(z_h) \leq h_i(z_h) + \epsilon\langle z_h\rangle^\gamma$. For those h and h_i observe that for every $z = (z_o, z_h^t)^t \in \mathcal{Z}$

$$\mathbb{1}\{z_o \leq h_i(z_h) - \epsilon\langle z_h\rangle^\gamma\} - \tau \leq \mathbb{1}\{z_o \leq h(z_h)\} - \tau \leq \mathbb{1}\{z_o \leq h_i(z_h) + \epsilon\langle z_h\rangle^\gamma\} - \tau.$$

From this follows that for every $h \in \mathcal{H}^{M_0}$, there exists h_i such that for every $(x, z) \in \mathcal{X} \times \mathcal{Z}$

$$\begin{aligned}
&K(x_1, x, b_n) \times \Big\{\mathbb{1}\{K \geq 0\}\mathbb{1}\{z_o \leq h_i(z_h) - \epsilon\langle z_h\rangle^\gamma\} \\
&\qquad + \mathbb{1}\{K < 0\}\mathbb{1}\{z_o \leq h_i(z_h) + \epsilon\langle z_h\rangle^\gamma\} - \tau\Big\} \\
\leq &K(x_1, x, b_n) \times \{\mathbb{1}\{z_o \leq h(z_h)\} - \tau\} \\
\leq &K(x_1, x, b_n) \times \Big\{\mathbb{1}\{K < 0\}\mathbb{1}\{z_o \leq h_i(z_h) - \epsilon\langle z_h\rangle^\gamma\} \\
&\qquad + \mathbb{1}\{K \geq 0\}\mathbb{1}\{z_o \leq h_i(z_h) + \epsilon\langle z_h\rangle^\gamma\} - \tau\Big\}
\end{aligned}$$

where $\mathbb{1}\{K < 0\}$ (*resp.* $\mathbb{1}\{K \geq 0\}$) is a shortcut for $\mathbb{1}\{K(x_1, x, b_n) < 0\}$ $\Big($*resp.* $\mathbb{1}\{K(x_1, x, b_n) \geq 0\}\Big)$.

We deduce that

$$\begin{aligned}
&\Big\{K(x_1, x, b_n) \times \Big\{\mathbb{1}\{K \geq 0\}\mathbb{1}\{z_o \leq h_i(z_h) - \epsilon\langle z_h\rangle^\gamma\} \\
&\qquad + \mathbb{1}\{K < 0\}\mathbb{1}\{z_o \leq h_i(z_h) + \epsilon\langle z_h\rangle^\gamma\} - \tau\Big\}, \\
&K(x_1, x, b_n) \times \Big\{\mathbb{1}\{K < 0\}\mathbb{1}\{z_o \leq h_i(z_h) - \epsilon\langle z_h\rangle^\gamma\} \\
&\qquad + \mathbb{1}\{K \geq 0\}\mathbb{1}\{z_o \leq h_i(z_h) + \epsilon\langle z_h\rangle^\gamma\} - \tau\Big\}\Big\}_{i=1}^{N(\epsilon)}
\end{aligned}$$

is a valid bracket of $\mathcal{F}_{n,x_1}$ with $L_2\left(P^{X,Z}\right)$ size

$$\begin{aligned}
&\sqrt{\mathbb{E}\left[K(x_1, X, b_n)^2 \times (\mathbb{1}\{Z_o \leq h_i(Z_h) + \epsilon\langle Z_h\rangle^\gamma\} - \mathbb{1}\{Z_o \leq h_i(Z_h) - \epsilon\langle Z_h\rangle^\gamma\})\right]} \\
=&\sqrt{\mathbb{E}\left[K(x_1, X, b_n)^2 \times \mathbb{E}\left[F_{Z_o|X,Z_h}(h_i(Z_h) + \epsilon\langle Z_h\rangle^\gamma) - F_{Z_o|X,Z_h}(h_i(Z_h) - \epsilon\langle Z_h\rangle^\gamma) \mid X\right]\right]} \\
\leq&\sqrt{2\epsilon \sup_{(z,x)\in\mathcal{Z}\times[0,1]^{d_x}} f_{Z_o|X,Z_h}(z_o \mid x, z_h)\mathbb{E}\left[K(x_1, X, b_n)^2 \langle Z_h\rangle^\gamma\right]}
\end{aligned}$$

where the last inequality is due to Assumption 3.7(ii).

This implies

$$N_{[]}\left(\epsilon\,||F_{n,x_1}||_{L_2(P^{X,Z})}, \mathcal{F}_{n,x_1}, L_2\left(P^{X,Z}\right)\right) \leq N\left(\epsilon^2 K(x_1, M_0, \gamma), \mathcal{H}^{M_0}, ||.||_{\infty,\gamma}\right),$$

where $K(x_1, M_0, \gamma) = \dfrac{||F_{n,x_1}||_{L_2(P^{X,Z})}}{\sqrt{2\sup_{(z,x)\in\mathcal{Z}\times[0,1]^{d_x}} f_{Z_o|X,Z_h}(z_o|x,z_h)\mathbb{E}\left[K(x_1,X,b_n)^2\langle Z_h\rangle^\gamma\right]}}$.

Since F_{n,x_1} is essentially bounded from below by $|K(x_1, x, b_n)|$ and $\sup_{x\in[0,1]^{d_x}} \mathbb{E}\left[\langle Z_h\rangle^\gamma \mid X = x\right] < +\infty$ under Assumption 3.7.(ii),

$$K(x_1, M_0, \gamma) \geq \left(2 \sup_{(z,x)\in\mathcal{Z}\times[0,1]^{d_x}} f_{Z_o|X,Z_h}(z_o \mid x, z_h) \sup_{x\in[0,1]^{d_x}} \mathbb{E}\left[\langle Z_h\rangle^\gamma \mid X = x\right]\right)^{-1/2} = K(\gamma)$$

which does not depend on x_1 anymore.

We conclude that uniformly in x_1

$$\sqrt{\log N_{[]}\left(\epsilon\,||F_{n,x_1}||_{L_2(P^{X,Z})}, \mathcal{F}_{n,x_1}, L_2\left(P^{\cdot|X_1=x_1}\right)\right)} \leq K(M_0)(\epsilon^2 K(\gamma))^{-d_{z_h}/2m}.$$

We can see than in the NPQIV model, whenever $m/d_{z_h} > 1$

$$\sup_{x_1 \in [0,1]^{d_x}} J(n, x_1, M_0, P) < +\infty.$$

Show that $\sup_{x_1 \in [0,1]^{d_x}} J_{l,l'}(n, x_1, M_0, P) < +\infty$ in the NPQIV case

As there is a single moment condition, we only have to consider the case $l = l' = 1$, *i.e* we focus on $\rho(z,h)^2 = (\mathbb{1}\{z_o \leq h(z_h)\} - \tau)^2$. We remark that

$$(\mathbb{1}\{z_o \leq h(z_h)\} - \tau)^2 = (1-2\tau)\mathbb{1}\{z_o \leq h(z_h)\} + \tau^2.$$

When $\tau = 1/2$, $\rho(z,h)^2 = \tau^2$ and the result is immediate. Otherwise, the result follows under exactly the same conditions as in the previous subsection. The only technicality arises when $\tau > 1/2$: in that case, $1 - 2\tau < 0$ and we have to exchange the roles of the upper and lower bracketing functions constructed in the previous subsection.

3.6.2.13 Proof of Lemma 3.14

This lemma is a consequence of Corollary 1 in [80]. We check that the conditions of that corollary are verified here for the class of functions $\mathcal{F} := \left\{\|m(\cdot,h)\|^2 : h \in \mathcal{H}^{M_0}\right\}$. Under Assumptions 3.2(ii) and 3.5(iii), remark that for every $(h_1, h_2) \in \mathcal{H}^{M_0} \times \mathcal{H}^{M_0}$ the reverse triangle inequality implies

$$\frac{1}{n}\sum_{i=1}^n \left|\|m(X_i,h_1)\|^2 - \|m(X_i,h_2)\|^2\right|^2 \leq C_1 \|h_1 - h_2\|^2_{\infty,\gamma},$$

where $C_1 := 4\sup_{(x,h)\in[0,1]^{d_x}\times\mathcal{H}^{M_0}} \|m(x,h)\|^2 L^2$.

This implies that for every $\epsilon > 0$

$$N\left(\epsilon, \mathcal{F}, L_2(P_n)\right) \leq N\left(\frac{\epsilon}{\sqrt{C_1}}, \mathcal{H}^{M_0}, \|\cdot\|_{\infty,\gamma}\right).$$

with $P_n := \frac{1}{n}\sum_{i=1}^n \delta_{\{X_i\}}$.

We explained in the proof of Lemma 3.13 that under Assumptions 3.3(i) and (iii) and $\gamma > m - d_{z_h}/2 > 0$, we have for every $\epsilon > 0$

$$N\left(\frac{\epsilon}{\sqrt{C_1}}, \mathcal{H}^{M_0}, \|\cdot\|_{\infty,\gamma}\right) \leq K(M_0)\exp\left\{\epsilon^{-d_{z_h}/m}\right\},$$

for some positive $K(M_0)$ that is finite for every $M_0 > 0$. Since C_1 can be chosen larger than 1 without loss of generality, we obtain $N\left(\epsilon/\sqrt{C_1}, \mathcal{F}, L_2(P_n)\right) \leq K(M_0)\exp\left\{(\epsilon/\sqrt{C_1})^{-d_{z_h}/m}\right\}$. By assumption, $m > d_{z_h}/2$ so that $d_{z_h}/m \in (0,2)$. What is more, remark that the constant C in Corollary 1 in [80] can be taken larger than 1 without loss of generality as well. As a result, we can apply this corollary with $A = K(M_0)$, $\alpha = d_{z_h}/m$, $q = 2$, $\epsilon = (0.5/C)^2$ and $\delta = (n\epsilon)^{-2/(\alpha+2)}$ to claim that for every $n \geq 1$

$$\begin{aligned}
&\mathbb{P}\left(\sup_{h\in\mathcal{H}^{M_0}:\mathbb{E}[\|m(X,h)\|^2]\geq 1.1\delta}\left|\frac{\frac{1}{n}\sum_{i=1}^n\|m(X_i,h)\|^2}{\mathbb{E}\left[\|m(X,h)\|^2\right]} - 1\right| > 0.5\right)\\
\leq&\mathbb{P}\left(\sup_{h\in\mathcal{H}^{M_0}:\mathbb{E}[\|m(X,h)\|^2]>\delta}\left|\frac{\frac{1}{n}\sum_{i=1}^n\|m(X_i,h)\|^2}{\mathbb{E}\left[\|m(X,h)\|^2\right]} - 1\right| > 0.5\right)\\
\leq&\frac{4}{3}\frac{4}{n\epsilon\delta}\exp\{-n\epsilon\delta/4\} = \frac{16}{3}\frac{1}{(\epsilon n)^{d_{z_h}/(d_{z_h}+2m)}}\exp\left\{-\frac{1}{4}(\epsilon n)^{d_{z_h}/(d_{z_h}+2m)}\right\}.
\end{aligned}$$

Pick $C_1 = 1.1\epsilon^{-2m/(d_{z_h}+2m)} > 1$, $C_2 = \frac{16}{3\epsilon^{d_{z_h}/(d_{z_h}+2m)}}$ and $C_3 = \frac{1}{4}\epsilon^{d_{z_h}/(d_{z_h}+2m)}$ to conclude.

Chapter 4

Empirical Process Results for Exchangeable Arrays

Abstract

Exchangeable arrays are natural ways to model common forms of dependence between units of a sample. Jointly exchangeable arrays are well suited to dyadic data, where observed random variables are indexed by two units from the same population. Examples include trade flows between countries or relationships in a network. Separately exchangeable arrays are well suited to multiway clustering, where units sharing the same cluster (e.g. geographical areas or sectors of activity when considering individual wages) may be dependent in an unrestricted way. We prove uniform laws of large numbers and central limit theorems for such exchangeable arrays. We obtain these results under the same moment restrictions and conditions on the class of functions as with i.i.d. data. As a result, convergence and asymptotic normality of nonlinear estimators can be obtained under the same regularity conditions as with i.i.d. data. We also show the convergence of bootstrap processes adapted to such arrays.

Keywords: exchangeable arrays, empirical processes, bootstrap.
Based on [51] : Davezies, L., D'Haultfœuille, X. & Guyonvarch Y., Empirical Process Results for Exchangeable Arrays. *Arxiv preprint*, arXiv:1906.11293, 2019.

4.1 Introduction

Taking into account dependence between observations is crucial for making correct inference. For instance, different observations may face common shocks, tending to correlate them positively and thus leading to overly optimistic inference when ignored [19]. A growing reason for the presence of such common shocks is that the data are polyadic (e.g., dyadic), namely they involve interactions between several units of a given population. An example is international trade, where each observation corresponds to a pair of countries, one exporting and the other importing. We can then expect that two such pairs may be dependent whenever they share at least one country, because of that country's specificities in terms of international trade. Another reason for common shocks is one-way or multiway clustering. In such cases, common shocks appear in one or several dimensions. For instance, wages of two individuals may be correlated either because they live in the same geographical area, or because they work in the

same sector.

[69] and [30] derived variance formulas for linear regressions with dyadic data and multiway clustering, respectively. The Stata command `ivreg2` and the R package `multiwaycov` are now used routinely to report standard errors accounting for multiway clustering. Perhaps surprisingly however, theory has lagged behind this practice. To our knowledge, the only paper showing the asymptotic validity of inference based on Fafchamps and Gubert's suggestion for dyadic data is [131]. Moreover, his result is restricted to OLS estimators only. Regarding multiway clustering, the only papers we are aware of are the recent works of [109] and [105].[1] Again, they focus on linear parameters.[2]

In this paper, we establish uniform laws of large numbers (LLN) and central limit theorems (CLT) for such type of data. Uniform LLNs and CLTs are key for showing consistency and asymptotic normality of nonlinear estimators under weak regularity conditions. As such, they have been studied extensively with i.i.d. but also dependent data. We refer to, e.g., [137] and [57] for overviews with respectively i.i.d. and time series data [see also, e.g., 18, for recent results on sampling designs]. Noteworthy, we obtain these uniform LLNs and CLTs under the same moment restrictions and conditions on the class of functions as with i.i.d. data. Thus, the results already obtained with i.i.d. data directly extend to the exchangeable arrays we consider. As a proof of concept, we consider such extensions for Z-estimators and smooth functionals of the empirical cumulative distribution function (cdf).

We also study consistency of the bootstrap. Specifically, we consider a direct generalization of the standard bootstrap for i.i.d. data to polyadic data. A related bootstrap scheme for multiway clustering is the so-called pigeonhole bootstrap, suggested by [108] and studied by [118], but for which no uniform result has been established so far. For both, we establish weak convergence of the corresponding process. These results imply the validity of the corresponding bootstrap schemes in a wide range of setting, including the Z-estimators and smooth functionals of the empirical cdf.

To prove these results, we first argue that polyadic data correspond to dissociated, jointly exchangeable arrays. Similarly, multiway clustering corresponds to dissociated separately exchangeable arrays. We then rely extensively on the so-called Aldous-Hoover-Kallenberg representation [89, 4, 95] for such arrays. This representation allows us in particular to prove a symmetrization lemma, which is very useful to derive the uniform LLNs and CLTs. This lemma generalizes a similar result for i.i.d. data, but also for U-processes [see, e.g. 56, Theorem 3.5.3]. Note that simple LLNs and CLTs have been already proved, or are direct consequences of known results on dissociated, jointly exchangeable arrays. For LLNs, we refer to [66] and Lemma 7.35 in [96]. For CLTs, see [128]. But to our knowledge, no abstract uniform LLNs and CLTs have been proved so far for such arrays. We therefore also contribute to this literature.

Finally, we illustrate our results with simulations and an application to international trade. A very popular model for explaining trade between countries is the so-called gravity equation, whose name is due to its similarities with the usual Newtonian gravity equation. Since [126], this equation has often been estimated with Poisson pseudo maximum likelihood, to deal in particular with the absence of trade between many countries. Our results apply to this nonlinear estimator. Using the same data and specification as [126], we show that much fewer explanatory variables are significant at usual levels when assuming dissociation and joint exchangeability rather than, e.g., i.i.d. observations [as in 126] or clustering along exporters or importers only, as is often done in the literature.

The paper is organized as follows. Section 4.2 describes the set-up and gives our main results. In addition to uniform LLNs and CLTs, we prove weak convergence of our bootstrap scheme. We also show

[1] See also our previous working paper [50], which is now superseded by this one.

[2] On the other hand and interestingly, [109] studies inference both with and without asymptotically normality. He also shows that refinements in asymptotic approximations are possible using the wild bootstrap.

results for Z-estimators and smooth functionals of the empirical cdf. In this section, we focus on jointly exchangeable arrays, as separately exchangeable arrays are more restrictive and thus can be essentially obtained as corollaries of these main results. In Section 4.3, we extend these findings to cases where the number of observations for each $k-$tuple (e.g., the number of matches between two sport players) varies. We also study separately exchangeable arrays. An important difference for such arrays is that the multiple dimensions, corresponding to different sources of clustering, may not grow at the same rate. We show that our results still hold in this case. Finally, the application to international trade is developed in Section 4.4.2. The proof of the symmetrization lemma is given in Appendix A. All other proofs are gathered in Appendix B.

4.2 The set up and main results

4.2.1 Set up

Before defining formally our data generating process, we introduce some notation. For any $A \subset \mathbb{R}$ and $B \subset \mathbb{R}^k$ for some $k \geq 2$, we let $A^+ = A \cap (0, +\infty)$ and

$$\overline{B} = \left\{ b = (b_1, ...b_k) \in B : \forall (i,j) \in \{1, ..., k\}^2, i \neq j, b_i \neq b_j \right\}.$$

We then let $\mathbb{I}_k = \overline{\mathbb{N}^{+k}}$ denote the set of k-tuples of $\mathbb{N}^+$ without repetition. Similarly, for any $n \in \mathbb{N}^+$, we let $\mathbb{I}_{n,k} = \overline{\{1, ..., n\}^k}$. For any $\boldsymbol{i} = (i_1, ..., i_k)$ and $\boldsymbol{j} = (j_1, ..., j_k)$ in $\mathbb{N}^k$, we let $\boldsymbol{i} \odot \boldsymbol{j} = (i_1 \times j_1, ..., i_k \times j_k)$. With a slight abuse of notation, we also let, for any $\boldsymbol{i} = (i_1, ..., i_k) \in \mathbb{N}^k$, $\{\boldsymbol{i}\}$ denote the set of distinct elements of $(i_1, ...i_k)$. For any $r \in \{1, ..., k\}$, we let

$$\mathcal{E}_r = \left\{ (e_1, ..., e_k) \in \{0,1\}^k : \sum_{j=1}^k e_j = r \right\}.$$

Finally, for any $A \subset \mathbb{N}^+$, we let $\mathfrak{S}(A)$ denote the set of permutations on A. For any $\boldsymbol{i} = (i_1, ..., i_k) \in \mathbb{N}^{+k}$ and $\pi \in \mathfrak{S}(\mathbb{N}^+)$, we let $\pi(\boldsymbol{i}) = (\pi(i_1), ..., \pi(i_k))$.

We are interested in polyadic data, that is to say random variables $Y_{\boldsymbol{i}}$ (whose support is denoted by $\mathcal{Y}$) indexed by $\boldsymbol{i} \in \mathbb{I}_k$. Dyadic data, which are the most common case, correspond to $k = 2$. For instance, when considering trade data, Y_{i_1,i_2} corresponds to export flows from country i_1 to country i_2. In network data, Y_{i_1,i_2} could be a dummy for whether there is a link from i_1 to i_2. In directed networks, $Y_{i_1,i_2} \neq Y_{i_2,i_1}$, while $Y_{i_1,i_2} = Y_{i_2,i_1}$ in undirected networks. Similarly, Y_{i_1,i_2,i_3} could capture whether (i_1, i_2, i_3) forms a triad or not [see, e.g. 139, for a motivation on triad counts]. $Y_{\boldsymbol{i}}$ could also correspond to data subject to multiway clustering. Then i_1,..., i_k are the indexes corresponding to the different dimensions of clustering, for instance geographical areas and sectors of activity. In such cases, however, adaptations of our set-up are needed, and we postpone this discussion to Section 4.3.2 below.

We assume that the random variables are generated according to a jointly exchangeable and dissociated array, defined formally as follows:

Assumption 4.1. *For any $\pi \in \mathfrak{S}(\mathbb{N}^+)$, $(Y_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{I}_k} \overset{d}{=} (Y_{\pi(\boldsymbol{i})})_{\boldsymbol{i} \in \mathbb{I}_k}$. Moreover, for any A, B disjoint subsets of $\mathbb{N}^+$ with $\min(|A|, |B|) \geq k$, $(Y_{\boldsymbol{i}})_{\boldsymbol{i} \in \overline{A^k}}$ is independent of $(Y_{\boldsymbol{i}})_{\boldsymbol{i} \in \overline{B^k}}$.*

The first part imposes that the labelling conveys no information: the joint distribution of the data remains identical under any possible permutation of the labels. The second part states that the array is dissociated: the variables are independent if they share no unit in common. For instance, $Y_{(i_1,i_2)}$ must be independent of $Y_{(j_1,j_2)}$ if $\{i_1, i_2\} \cap \{j_1, j_2\} = \emptyset$. On the other hand, Assumption 4.1 does not impose independence

otherwise. This is important in many applications. In the international trade example, Y_{i_1,i_2} and Y_{i_1,i_3} are likely to be dependent because if i_1 is open to international trade, it tends to export more than the average to any other country. It may also import more from other countries, meaning that Y_{i_1,i_2} and Y_{i_3,i_1} could also be dependent.

Lemma 4.1 below is very helpful to better understand the dependence structure imposed by joint exchangeability and dissociation. It may be seen as an extension of de Finetti's theorem to arrays satisfying such restrictions. It is also key for establishing our asymptotic results below.

Lemma 4.1. *Assumption 4.1 holds if and only if there exist i.i.d. variables* $(U_J)_{J\subset\mathbb{N}^+,1\leq|J|\leq k}$ *and a measurable function* τ *such that almost surely,*[3]

$$Y_{\boldsymbol{i}} = \tau\left((U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{\boldsymbol{e}\in\cup_{r=1}^k\mathcal{E}_r}\right) \quad \forall \boldsymbol{i}\in\mathbb{I}_k. \tag{4.1}$$

This result is due to [95] but a weaker version, where the equality only holds in distribution, is known as Aldous-Hoover representation [4, 89]. Accordingly, we refer to (4.1) as the AHK representation hereafter. To illustrate it, let us consider dyadic data ($k=2$). Then, according to Lemma 4.1, we have, for every $i_1<i_2$,

$$Y_{i_1,i_2} = \tau(U_{i_1}, U_{i_2}, U_{\{i_1,i_2\}}). \tag{4.2}$$

Thus, in the example of trade flows, the volume of exports from i_1 to i_2 depends on factors specific to i_1 and i_2, such as their own GDP, but also on factors relating both, such as the distance between the two countries. Note also the link between (4.2) and U-statistics: Y_{i_1,i_2} would correspond to such a statistic if τ did not depend on its third argument.

Under Assumption 4.1, the $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$ have a common marginal probability distribution, which we denote by P. We are interested in estimating and making inference on features of this distribution, such as its expectation or a quantile, based on observing the first n units only, namely the sample $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_{n,k}}$, with $n\geq k$.

4.2.2 Uniform laws of large numbers and central limit theorems

Let $\mathcal{F}$ denote a class of real-valued functions admitting a first moment with respect to the distribution P and let Pf denote the corresponding moment $\mathbb{E}[f(Y_{\mathbf{1}})]$. To avoid measurability issues and the use of outer expectations subsequently, we maintain the following assumption:

Assumption 4.2. *There exists a countable subclass* $\mathcal{G}\subset\mathcal{F}$ *such that elements of* $\mathcal{F}$ *are pointwise limits of elements of* $\mathcal{G}$*.*

Assumption 4.2 is not necessary but often imposed [see, e.g. 43, 98]. We refer to Kosorok (2006, pp.137-140) for further discussion.

In this section, we study the empirical measure $\mathbb{P}_n$ and the empirical process $\mathbb{G}_n$ defined on $\mathcal{F}$ by

$$\mathbb{P}_n f = \frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} f(Y_{\boldsymbol{i}}),$$

$$\mathbb{G}_n f = \sqrt{n}\left(\mathbb{P}_n f - Pf\right).$$

We prove below that under restrictions on $\mathcal{F}$, $\mathbb{P}_n f$ converges almost surely to Pf uniformly over $f\in\mathcal{F}$, while $\mathbb{G}_n$ converges weakly to a Gaussian process as n tends to infinity. We refer to, e.g., [137] for a

[3] In this formula, the $(U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{\boldsymbol{e}\in\cup_{r=1}^k\mathcal{E}_r}$ appear according to a precise ordering, which we let nonetheless implicit as it bears no importance hereafter.

formal definition of weak convergence of empirical processes. These results, which are stronger than pointwise convergence of $\mathbb{P}_n f$ and $\mathbb{G}_n f$, are key in establishing the consistency and asymptotic normality of, e.g., smooth functionals of the empirical cdf or Z- and M-estimators. We consider briefly applications in Section 4.2.4 below, and refer to Part 3 of [137] for a more comprehensive review of statistical applications of empirical process results.

We use the rate $\sqrt{n}$ to normalize $\mathbb{P}_n f - Pf$, though we have $n!/(n-k)!$ different random variables. In general, we cannot expect a better rate of convergence. To see this, let $(X_i)_{i\in\mathbb{N}^+}$ be i.i.d. random variables and let $Y_{\boldsymbol{i}} = \sum_{j\in\{\boldsymbol{i}\}} X_j$. Then $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$ satisfies Assumption 4.1, and $\mathbb{P}_n f$ boils down to an average over n i.i.d. terms only. In some cases, however, for instance if the $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$ are i.i.d., the convergence rate is faster than $\sqrt{n}$. Theorem 4.1 below remains valid in such cases, but the limit Gaussian process is then degenerate.

To establish uniform LLNs and CLTs with i.i.d. data $(X_i)_{i\in\mathbb{N}^+}$, a natural way to proceed is to show a symmetrization lemma [see, e.g., Lemma 2.3.1 in 137]. Such a lemma states that for any non-decreasing convex function Φ from $\mathbb{R}^+$ to $\mathbb{R}$ and i.i.d. Rademacher variables $(\varepsilon_1, ..., \varepsilon_n)$ independent of $(X_1, ..., X_n)$,

$$\mathbb{E}\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - E(f(X_1)\right|\right)\right] \leq \mathbb{E}\left[\Phi\left(2\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right)\right]. \tag{4.3}$$

This inequality is useful for proving uniform LLNs and CLTs because conditional on $(X_1, ..., X_n)$, the process $f \mapsto \sum_{i=1}^n \varepsilon_i f(X_i)$ is subgaussian, implying that we can apply maximal inequalities to it. Our main insight is that (4.3) generalizes to jointly exchangeable and dissociated arrays. Let $(\varepsilon_A)_{A\subset\mathbb{N}^+}$ denote Rademacher independent variables, independent of $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$. Then:

Lemma 4.2. *If Assumptions 4.1-4.2 hold and $P|f| < +\infty$ for all $f \in \mathcal{F}$, there exist real numbers $C_{1,k}, ..., C_{k,k}$ depending only on k and $(Y^1_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$,..., $(Y^k_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$ jointly exchangeable and dissociated arrays with marginal distribution P such that*

$$\begin{aligned}&\mathbb{E}\left[\Phi\left(\sup_{F\in\mathcal{F}}|\mathbb{P}_n f - Pf|\right)\right]\\ \leq&\frac{1}{k}\sum_{r=1}^{k}\frac{1}{|\mathcal{E}_r|}\sum_{\boldsymbol{e}'\in\mathcal{E}_r}\mathbb{E}\left[\Phi\left(\frac{(n-k)!}{n!}C_{r,k}\sup_{f\in\mathcal{F}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+} f\left(Y^r_{\boldsymbol{i}}\right)\right|\right)\right],\end{aligned}$$

Though the inequality is more complicated than (4.3), it serves the exact same purpose as with i.i.d. data: conditional on the $(Y^r_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$, the process $f \mapsto \sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} \varepsilon_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+} f\left(Y^r_{\boldsymbol{i}}\right)$ is still subgaussian. In view of the AHK representation, the terms $\varepsilon_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}$ could be expected. Given the aforementioned link with U-statistics, Lemma 4.2 can also be seen as a generalization of the symmetrization lemma for U-processes, see in particular Theorem 3.5.3 in [56].

The proof of Lemma 4.2, given in the appendix (Section 4.6.1), relies extensively on Lemma 4.1 and a decoupling inequality that may be of independent interest (see Lemma 4.3). The latter result generalizes a similar inequality for U-processes [see 55]. In the proofs of both lemmas, we follow similar strategies as with U-processes, with two complications. First, even with $k = 2$, $Y_{\boldsymbol{i}}$ does not only depend on U_{i_1} and U_{i_2}, but also on $U_{\{i_1,i_2\}}$. Second, when $k \geq 3$, dependence between observations arises not only because of single-unit terms such as U_{i_1} or U_{i_2}, but also because of multiple-unit terms such as $U_{\{i_1,i_2\}}$. Related to that, it is unclear to us whether one can always replace (up to adjusting $C_{r,k}$) $Y^r_{\boldsymbol{i}}$ by $Y_{\boldsymbol{i}}$ in Lemma 4.2. Such a result holds true for $k \leq 3$, using a reverse decoupling inequality, but this inequality may not be valid for all (r, k). See Appendix A (Section 4.6.1.1) for more details on the matter.

Lemma 4.2 allows us to extend the uniform LLNs and CLTs for i.i.d. data to jointly exchangeable and dissociated arrays, under the same restrictions on the class $\mathcal{F}$. Subsequently, an envelope of $\mathcal{F}$ is a

measurable function F satisfying $F(u) \geq \sup_{f\in\mathcal{F}} |f(u)|$. For any $\eta > 0$ and any seminorm $||.||$ on a space containing $\mathcal{F}$, $N(\eta, \mathcal{F}, ||.||)$ denotes the minimal number of $||.||$-closed balls of radius η with centers in $\mathcal{F}$ needed to cover $\mathcal{F}$. The seminorms we consider hereafter are $\|f\|_{\mu,r} = (\int |f|^r d\mu)^{1/r}$ for any $r \geq 1$ and probability measure or cdf function μ.

Assumption 4.3. *The class $\mathcal{F}$ admits an envelope F with*

$$\forall \eta > 0, \sup_Q N\left(\eta ||F||_{Q,1}, \mathcal{F}, ||.||_{Q,1}\right) < \infty,$$

where the supremum is taken over the set of probability measures with finite support on $\mathcal{Y}$.

Assumption 4.4. *The class $\mathcal{F}$ admits an envelope F with*

$$\int_0^{+\infty} \sup_Q \sqrt{\log N\left(\eta ||F||_{Q,2}, \mathcal{F}, ||.||_{Q,2}\right)} d\eta < +\infty,$$

where the supremum is taken over the set of probability measures with finite support on $\mathcal{Y}$.

Assumptions 4.3 and 4.4 are exactly the same as the conditions often imposed with i.i.d. data to show uniform LLNs and CLTs [see, e.g., Theorems 19.13 and 19.14 in 136].[4] In particular, Assumption 4.4 imposes a condition on what is usually referred to as the uniform entropy integral, see, e.g., [137]. Finiteness of the uniform entropy integral is satisfied by any VC-type class of functions [see 43, for a definition], or by the convex hull of such classes under some restrictions. The following theorem establishes uniform LLNs and CLTs under these two conditions. As of now, we denote by $\mathbf{1}$ and $\mathbf{1}'$ the $k-$tuples $(1, ..., k)$ and $(1, k+1, ..., 2k-1)$, respectively.

Theorem 4.1. *Suppose that Assumptions 4.1-4.2 hold. Then:*

1. *If Assumption 4.3 holds with F also satisfying $PF < +\infty$, $\sup_{f\in\mathcal{F}} |\mathbb{P}_n f - Pf|$ tends to 0 almost surely and in L^1.*
2. *If Assumption 4.4 holds with F also satisfying $PF^2 < +\infty$, the process $\mathbb{G}_n$ converges weakly to a centered Gaussian process $\mathbb{G}$ on $\mathcal{F}$ as n tends to infinity. Moreover, the covariance kernel K of $\mathbb{G}$ satisfies:*

$$K(f_1, f_2) = \frac{1}{(k-1)!^2} \sum_{(\pi,\pi')\in\mathfrak{S}(\{\mathbf{1}\})\times\mathfrak{S}(\{\mathbf{1}'\})} \mathbb{C}ov\left(f_1(Y_{\pi(\mathbf{1})}), f_2(Y_{\pi'(\mathbf{1}')})\right).$$

When $\mathcal{F}$ is finite, Part 1 can be proved by combining Theorem 3 in [66] and Lemma 7.35 in [96]. The result for an infinite class, however, does not follow from these results, whereas it does follow from Lemma 4.2 coupled with standard tools from empirical process theory. Similarly, Part 2 was proved for a finite $\mathcal{F}$ by [128]. However, the asymptotic equicontinuity of $\mathbb{G}_n$, which is necessary when $\mathcal{F}$ is infinite, is difficult to prove. Again Lemma 4.2 is a core ingredient in this respect.

4.2.3 Convergence of the bootstrap process

In this section, we study the properties of the following bootstrap sampling scheme:

1. n units are sampled independently in $\{1, ..., n\}$ with replacement and equal probability. W_i denotes the number of times unit i is sampled.

[4] In [136], the supremum in Assumptions 4.3 and 4.4 is taken over the set of probability measures Q with finite support on $\mathcal{Y}$ and such that $||F||_{Q,2} > 0$. This additional restriction is simply due to a different convention in constructing covering numbers, as van der Vaart considers open balls while we use closed balls, following, e.g., [98].

2. the $k-$tuple $\boldsymbol{i} = (i_1, ..., i_k) \in \mathbb{I}_{n,k}$ is then selected $W_{\boldsymbol{i}} = \prod_{j=1}^{k} W_{i_j}$ times in the bootstrap sample.

Then we consider $\mathbb{P}_n^*$ and $\mathbb{G}_n^*$, defined on $\mathcal{F}$ by

$$\mathbb{P}_n^* f = \frac{(n-k)!}{n!} \sum_{\boldsymbol{i} \in \mathbb{I}_{n,k}} W_{\boldsymbol{i}} f(Y_{\boldsymbol{i}}),$$

$$\mathbb{G}_n^* f = \sqrt{n} \left(\mathbb{P}_n^* f - \mathbb{P}_n f \right).$$

Asymptotic validity of the bootstrap amounts to showing that conditional on the data $(Y_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{I}_k}$, $\mathbb{G}_n^*$ converges weakly to the process $\mathbb{G}$ defined in Theorem 4.1. As discussed in, e.g., van der Vaart and Wellner (1996, Chapter 3.6), the almost-sure conditional weak convergence boils down to proving

$$\sup_{h \in \mathsf{BL}_1} \left| \mathbb{E}\left(h(\mathbb{G}_n^*) \middle| (Y_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{I}_k} \right) - \mathbb{E}\left(h(\mathbb{G}) \right) \right| \xrightarrow{a.s.} 0, \tag{4.4}$$

where BL_1 is the set of bounded and Lipschitz functions from $\ell^\infty(\mathcal{F})$ to $[0,1]$.

Theorem 4.2. *Suppose that Assumptions 4.1-4.2 and 4.4 hold, with F also satisfying $PF^2 < +\infty$. Then, conditional on $(Y_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{I}_k}$ and almost surely, the process $\mathbb{G}_n^*$ converges weakly to $\mathbb{G}$.*

This theorem ensures the asymptotic validity of the bootstrap above not only for sample means, but also for smooth functionals of the empirical cdf and nonlinear estimators, as we shall see below. The proof of Theorem 4.2 follows the same lines as that of Theorem 4.1, though some of the corresponding steps are more involved, as often with the bootstrap. In particular, to prove pointwise convergence, we use arguments in Lindeberg's proof of the CLT for triangular arrays, Theorem 4.1.1 and Urysohn's subsequence principle, combined with Prohorov's theorem.

Note that in contrast with the standard bootstrap for i.i.d. data,

$$\mathbb{E}\left(\mathbb{P}_n^*(f) \middle| (Y_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{I}_k} \right) = \frac{1}{n^k} \sum_{\boldsymbol{i} \in \mathbb{I}_{n,k}} f(Y_{\boldsymbol{i}}) \neq \mathbb{P}_n f.$$

However, the difference between $\mathbb{P}_n$ and $\mathbb{P}'_n$, the empirical measure with weights $1/n^k$, becomes negligible as $n \to \infty$. Accordingly, we also show in the proof of Theorem 4.2 the almost-sure conditional convergence of $\sqrt{n}\left(\mathbb{P}_n^* f - \mathbb{P}'_n f \right)$, in addition to that of $\mathbb{G}_n^*$.

4.2.4 Application to nonlinear estimators

Theorem 4.1 ensures the root-n consistency and asymptotic normality of a large class of estimators. In turn, Theorem 4.2 shows that using the bootstrap for such estimators is asymptotically valid. To illustrate these points, we consider here two popular classes of estimators, namely Z-estimators and smooth functionals of the empirical cdf. Similar results could be obtained for, e.g., M- or GMM estimators.

Let us first consider Z-estimators. Let Θ denote a normed space, endowed with the norm $\|.\|_\Theta$ and let $(\psi_{\theta,h})_{(\theta,h) \in \Theta \times \mathcal{H}}$ denote a class of real, measurable functions. Let $\Psi(\theta)(h) = P\psi_{\theta,h}$, $\Psi_n(\theta)(h) = \mathbb{P}_n \psi_{\theta,h}$ and $\Psi_n^*(\theta)(h) = \mathbb{P}_n^* \psi_{\theta,h}$. We let, for any real function g on $\mathcal{H}$, $\|g\|_\mathcal{H} = \sup_{h \in \mathcal{H}} |g(h)|$. The parameter of interest θ_0, which satisfies $\Psi(\theta_0) = 0$, is estimated by $\widehat{\theta}_n = \arg\min_{\theta \in \Theta} \|\Psi_n(\theta)\|_\mathcal{H}$. We also define $\widehat{\theta}_n^* = \arg\min_{\theta \in \Theta} \|\mathbb{P}_n^* \psi_{\theta,h}\|_\mathcal{H}$ as the bootstrap counterpart of $\widehat{\theta}_n$. The following theorem extends Theorem 13.4 in [102] to jointly exchangeable and dissociated arrays.

Theorem 4.3. *Suppose that Assumption 4.1 holds and:*

1. *$\|\Psi(\theta_m)\|_\mathcal{H} \to 0$ implies $\|\theta_m - \theta_0\|_\Theta \to 0$ for every $(\theta_m)_{m \in \mathbb{N}}$ in Θ;*

2. *The class* $\{\psi_{\theta,h} : (\theta, h) \in \Theta \times \mathcal{H}\}$ *satisfies Assumptions 4.2-4.3, with the envelope function* F *satisfying* $PF < +\infty$*;*

3. *There exists* $\delta > 0$ *such that the class* $\{\psi_{\theta,h} : \|\theta - \theta_0\|_\Theta < \delta, h \in \mathcal{H}\}$ *satisfies Assumptions 4.2 and 4.4, with an envelope function* F_δ *satisfying* $PF_\delta^2 < +\infty$*;*

4. $\lim_{\theta\to\theta_0} \sup_{h\in\mathcal{H}} P\left(\psi_{\theta,h} - \psi_{\theta_0,h}\right)^2 = 0$*;*

5. $\|\Psi_n(\widehat{\theta})\|_\mathcal{H} = o_p(n^{-1/2})$ *and* $P\left(\|\sqrt{n}\Psi_n^*(\widehat{\theta}^*)\|_\mathcal{H} > \eta | (Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right) = o_p(1)$ *for every* $\eta > 0$*;*

6. $\theta \mapsto \Psi(\theta)$ *is Fréchet-differentiable at* θ_0*, with continuously invertible derivative* $\dot{\Psi}_{\theta_0}$*.*

Then $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ *converges in distribution to a centered Gaussian process* $\mathbb{G}$*. Moreover, conditional on* $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$ *and almost surely,* $\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta})$ *converges in distribution to* $\mathbb{G}$*.*

Next, we consider smooth functionals of F_Y, the cdf of $Y_{\boldsymbol{i}}$. Specifically, suppose that $\mathcal{Y} \subset \mathbb{R}^p$ for some $p \in \mathbb{N}^+$ and $\theta_0 = g(F_Y)$, where g is Hadamard differentiable [for a definition, see, e.g., 137, Section 3.9.1]. We estimate θ_0 with $\widehat{\theta} = g(\widehat{F_Y})$, where $\widehat{F_Y}$ denotes the empirical cdf of $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_{n,k}}$. Finally, we let $\widehat{\theta}^*$ denote the bootstrap counterpart of $\widehat{\theta}$.

Theorem 4.4. *Suppose that* g *is Hadamard differentiable at* F_Y *tangentially to a set* $\mathbb{D}_0$*, with derivative equal to* g'_{F_Y}*. Suppose also that Assumption 4.1 holds. Then:*

1. $\sqrt{n}(\widehat{F_Y} - F_Y)$ *converges weakly, as a process indexed by* y*, to a Gaussian process* $\mathbb{G}$ *with kernel* K *satisfying*

$$K(y_1, y_2) = \frac{1}{(k-1)!^2} \sum_{(\pi,\pi')\in\mathfrak{S}(\{\mathbf{1}\})\times\mathfrak{S}(\{\mathbf{1}'\})} \mathbb{C}ov\big(\mathbb{1}_{\{Y_{\pi(\mathbf{1})}\le y_1\}}, \mathbb{1}_{\{Y_{\pi'(\mathbf{1}')}\le y_2\}}\big).$$

2. *If* $\mathbb{G} \in \mathbb{D}_0$ *with probability one,*

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow[n\to+\infty]{d} \mathcal{N}(0, \mathbb{V}(g'_{F_Y}(\mathbb{G}))).$$

Moreover, conditional on $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$ *and almost surely,* $\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_0)$ *converges in distribution to the same limit.*

In practice, $\mathbb{D}_0$ often corresponds to the set of functions that are continuous everywhere or at a certain point y_0. This is the case for instance with $g : F_Y \mapsto F_Y^{-1}(\tau)$ for $\tau \in (0, 1)$. In such cases, one can show that $\mathbb{G} \in \mathbb{D}_0$ under the same condition as for i.i.d. data, namely that F_Y is continuous everywhere or at the point $F_Y^{-1}(\tau)$.

4.3 Extensions

4.3.1 Heterogeneous number of observations

In some cases, we observe multiple observations for the same $k-$tuple $\boldsymbol{i}$. For instance, in the case of exchanges in a network, we may observe multiple or no such exchanges between i_1 and i_2. In sport competitions, we may observe N_{i_1,i_2} matches between players i_1 and i_2, with possibly $N_{i_1,i_2} = 0$. To deal with this issue, we consider that for each $\boldsymbol{i} \in \mathbb{I}_k$, there exists a random variable $N_{\boldsymbol{i}} \in \mathbb{N}$ and a sequence $Y_{\boldsymbol{i}} = (Y_{\boldsymbol{i},\ell})_{\ell\ge1}$, with $Y_{\boldsymbol{i},\ell}$ having support $\mathcal{Y}$, such that we only observe $(N_{\boldsymbol{i}}, (Y_{\boldsymbol{i},\ell})_{1\le\ell\le N_{\boldsymbol{i}}})$. To allow for $N_{\boldsymbol{i}} = 0$, we assume in the following that for any sequence $(a_\ell)_{\ell\ge1}$, $\sum_{\ell=1}^0 a_\ell = 0$.

In this set-up, it is often natural to redefine the parameters of interest: if the relevant units of observation are the $N_{\boldsymbol{i}}$ units within each $k-$tuple, then parameters of interest are defined with respect to $\widetilde{P}$ rather than P, with

$$\widetilde{P}f = \mathbb{E}\left[\sum_{\ell=1}^{N_{\mathbf{1}}} f(Y_{\mathbf{1},\ell})\right].$$

In the example of sport matches, this expectation weights equally each match rather than each pair of players and is therefore often more relevant. For instance, the sample average

$$\widehat{\theta} = \frac{\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} \sum_{\ell=1}^{N_{\boldsymbol{i}}} Y_{\boldsymbol{i},\ell}}{\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} N_{\boldsymbol{i}}}$$

is an estimator of $\theta_0 = \widetilde{P}(\mathsf{Id})/\widetilde{P}(1)$, where Id denotes the identity function. This parameter also satisfies $\theta_0 = \int y d\widetilde{F}_Y(y)$, with $\widetilde{F}_Y(y) = \widetilde{P}(\mathbb{1}_{\{\cdot\leq y\}})/\widetilde{P}(1)$. Similarly, quantiles would be defined as $\theta_0 = \widetilde{F}_Y^{-1}(\tau)$ for some $\tau \in (0,1)$. More generally, any parameter related to the units within each $k-$tuple is defined with respect to $\widetilde{P}$ rather than P.

Accordingly, we study the behavior of $\widetilde{\mathbb{P}}_n$, $\widetilde{\mathbb{G}}_n$ and $\widetilde{\mathbb{G}}_n^*$ defined on $\mathcal{F}$ by:

$$\widetilde{\mathbb{P}}_n f = \frac{(n-k)!}{n!} \sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} \sum_{\ell=1}^{N_{\boldsymbol{i}}} f(Y_{\boldsymbol{i},\ell}),$$

$$\widetilde{\mathbb{G}}_n f = \sqrt{n}\left(\widetilde{\mathbb{P}}_n(f) - \widetilde{P}f\right),$$

$$\widetilde{\mathbb{G}}_n^* f = \sqrt{n}\frac{(n-k)!}{n!} \sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} (W_{\boldsymbol{i}} - 1) \sum_{\ell=1}^{N_{\boldsymbol{i}}} f(Y_{\boldsymbol{i},\ell}).$$

The following theorem shows that the previous results extend to this set-up with random $N_{\boldsymbol{i}}$, only up to adaptations of the moment conditions.

Theorem 4.5. *Suppose that Assumption 4.1 holds with $(N_{\boldsymbol{i}}, Y_{\boldsymbol{i}})$ in place of $Y_{\boldsymbol{i}}$, $\widetilde{P}1 > 0$ and Assumption 4.2 holds. Then:*

1. *If Assumption 4.3 holds with F also satisfying $\widetilde{P}F < +\infty$, then $\sup_{f\in\mathcal{F}} \left|\widetilde{\mathbb{P}}_n f - \widetilde{P}f\right|$ tends to 0 almost surely and in L^1.*
2. *If $\mathbb{E}(N_{\mathbf{1}}^2) < +\infty$ and Assumption 4.4 holds with F also satisfying $\mathbb{E}\left(N_{\mathbf{1}} \sum_{\ell=1}^{N_{\mathbf{1}}} F^2(Y_{\mathbf{1},\ell})\right) < +\infty$, the process $\widetilde{\mathbb{G}}_n$ converges weakly to a centered Gaussian process $\widetilde{\mathbb{G}}$ on $\mathcal{F}$ as n tends to infinity. Moreover, the covariance kernel $\widetilde{K}$ of $\widetilde{\mathbb{G}}$ satisfies:*

$$\widetilde{K}(f_1, f_2) = \frac{1}{(k-1)!^2} \sum_{(\pi,\pi')\in\mathfrak{S}(\{\mathbf{1}\})\times\mathfrak{S}(\{\mathbf{1}'\})} \mathbb{C}ov\left(\sum_{\ell=1}^{N_{\pi(\mathbf{1})}} f_1(Y_{\pi(\mathbf{1}),\ell}), \sum_{\ell=1}^{N_{\pi'(\mathbf{1}')}} f_2(Y_{\pi'(\mathbf{1}'),\ell})\right).$$

3. *Under the same condition as in 2., the process $\widetilde{\mathbb{G}}_n^*$ converges weakly to $\widetilde{\mathbb{G}}$, conditional on $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$ and almost surely.*

We assume that $(N_{\boldsymbol{i}}, Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$, rather than just $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$, satisfies Assumption 4.1. Importantly, however, this does not restrict the dependence between $N_{\boldsymbol{i}}$ and $Y_{\boldsymbol{i}}$, or between the $(Y_{\boldsymbol{i},\ell})_\ell$. Hence, conditional on $N_{\boldsymbol{i}}$, the correlation between $Y_{\boldsymbol{i},\ell}$ and $Y_{\boldsymbol{i},\ell'}$ may vary with $N_{\boldsymbol{i}}$, for instance. Note also that even if we focus on $\widetilde{P}$ rather than P here, the conditions on $\mathcal{F}$ remain nearly unchanged, with only modifications of the moment conditions. For uniform LLNs, we simply replace $PF < +\infty$ by $\widetilde{P}F < +\infty$. For uniform CLTs, instead of replacing $PF^2 < +\infty$ by $\widetilde{P}F^2 < +\infty$, we require the slightly stronger conditions

that $\mathbb{E}(N_{\mathbf{1}}^2) < +\infty$ and $\mathbb{E}\left(N_{\mathbf{1}} \sum_{\ell=1}^{N_{\mathbf{1}}} F^2(Y_{\mathbf{1},\ell})\right) < +\infty$. These conditions are nonetheless equivalent to $\widetilde{P}F^2 < +\infty$ when $N_{\mathbf{1}}$ is bounded. Note also that with a finite $\mathcal{F}$, our proof would only require $\widetilde{P}F^2 < +\infty$.

The proof of Theorem 4.5 is very similar to those of Theorems 4.1 and 4.2, with one difference. In those theorems, we use the symmetrization lemma to bound the fluctuations of $\mathbb{G}_n$ by a function of the entropy of the class $\mathcal{F}$. Here, similarly, we bound the fluctuations of $\widetilde{\mathbb{G}}_n$ by a function of the entropy of the class

$$\widetilde{\mathcal{F}} = \left\{ \widetilde{f}(n, y_1, ..., y_n) = \sum_{\ell=1}^{n} f(y_\ell) : n \in \mathbb{N}, (y_1, ..., y_n) \in \mathcal{Y}^n; f \in \mathcal{F} \right\}.$$

The additional point to prove is that we can control the complexity of $\widetilde{\mathcal{F}}$ under Assumption 4.4 and the moment conditions above, even if Assumption 4.4 imposes conditions on $\mathcal{F}$ rather than on $\widetilde{\mathcal{F}}$ directly.

4.3.2 Separately exchangeable arrays

Up to now, we have considered cases where the n units that interact stem from the same population. In some cases, however, they do not, because the k populations differ. For instance, we may be interested only in relationships between men and women. In that case, the symmetry condition in Assumption 4.1 has to be strengthened: both the labelling of men and the labelling of women should be irrelevant. This corresponds to so-called separately exchangeable arrays, defined formally in Assumption 4.5 below. Another important motivation for considering separately exchangeable arrays is multiway clustering, namely dependence arising through different dimensions of clustering. For instance, wages of workers may be affected by local shocks or sector-of-activity shocks. In such cases, we observe $Y_{i_1,i_2,\ell}$, the wage of worker ℓ in geographical area i_1 and sector of activity i_2.

More generally, we consider in this section random variables $Y_{\boldsymbol{i}} = (Y_{\boldsymbol{i},\ell})_{\ell \geq 1}$, where $\boldsymbol{i} = (i_1, ..., i_k) \in \mathbb{N}^{+k}$, implying that repetitions (e.g. $\boldsymbol{i} = (1, ..., 1)$) are allowed. As above, we only observe, for each $k-$tuple $\boldsymbol{i}$, $(Y_{\boldsymbol{i},1}, ..., Y_{\boldsymbol{i},N_{\boldsymbol{i}}})$. We impose the following condition on these random variables.

Assumption 4.5. *For any* $(\pi_1, ..., \pi_k) \in \mathfrak{S}(\mathbb{N}^+)^k$*,*

$$(N_{\boldsymbol{i}}, Y_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{N}^{+k}} \overset{d}{=} (N_{\pi_1(i_1),...,\pi_k(i_k)}, Y_{\pi_1(i_1),...,\pi_k(i_k)})_{\boldsymbol{i} \in \mathbb{N}^{+k}}.$$

Moreover, for any A, B*, disjoint subsets of* $\mathbb{N}^+$*,* $(N_{\boldsymbol{i}}, Y_{\boldsymbol{i}})_{\boldsymbol{i} \in A^k}$ *is independent of* $(N_{\boldsymbol{i}}, Y_{\boldsymbol{i}})_{\boldsymbol{i} \in B^k}$*.*

This condition is stronger than Assumption 4.1 since it implies in particular equality in distribution for $\pi_1 = ... = \pi_k$.

Let us redefine $\mathbf{1}$ here as $(1, ..., 1)$ and let $\boldsymbol{n} = (n_1, ..., n_k)$, where $n_j \geq 1$ denotes the number of units observed in population or cluster j. Note that in general, $n_j \neq n_{j'}$ for $j \neq j'$. The sample at hand is then $(N_{\boldsymbol{i}}, (Y_{\boldsymbol{i},\ell})_{1 \leq \ell \leq N_{\boldsymbol{i}}})_{\mathbf{1} \leq \boldsymbol{i} \leq \boldsymbol{n}}$, where $\boldsymbol{i} \geq \boldsymbol{i}'$ means that $i_j \geq i'_j$ for all $j = 1, ..., k$. Let $\underline{n} = \min(n_1, ..., n_k)$. The empirical measure and empirical process that we consider for separately exchangeable arrays are:

$$\widetilde{\mathbb{P}}_{\boldsymbol{n}} f = \frac{1}{\prod_{j=1}^{k} n_j} \sum_{\mathbf{1} \leq \boldsymbol{i} \leq \boldsymbol{n}} \sum_{\ell=1}^{N_{\boldsymbol{i}}} f(Y_{\boldsymbol{i},\ell}),$$
$$\widetilde{\mathbb{G}}_{\boldsymbol{n}} f = \sqrt{\underline{n}} \left(\widetilde{\mathbb{P}}_{\boldsymbol{n}} f - \widetilde{P} f \right).$$

We also consider the "pigeonhole bootstrap", suggested by [108] and studied, in the case of the sample mean and for particular models, by [118]. This bootstrap scheme is very close to the one we considered in Section 4.2 for jointly exchangeable arrays, except that the weights are now independent from one coordinate to another:

1. For each $j \in \{1, ..., k\}$, n_j elements are sampled with replacement and equal probability in the set $\{1, ..., n_j\}$. For each i_j in this set, let $W_{i_j}^j$ denote the number of times i_j is selected this way.
2. The k-tuple $\boldsymbol{i} = (i_1, ..., i_k)$ is then selected $W_{\boldsymbol{i}} = \prod_{j=1}^k W_{i_j}^j$ times in the bootstrap sample.

The bootstrap process $\widetilde{\mathbb{G}}_{\boldsymbol{n}}^*$ is thus defined on $\mathcal{F}$ by

$$\widetilde{\mathbb{G}}_{\boldsymbol{n}}^* f = \sqrt{\underline{n}} \left(\frac{1}{\prod_{j=1}^k n_j} \sum_{\mathbf{1} \leq \boldsymbol{i} \leq \boldsymbol{n}} (W_{\boldsymbol{i}} - 1) \sum_{\ell=1}^{N_{\boldsymbol{i}}} f(Y_{\boldsymbol{i},\ell}) \right).$$

Henceforth, we consider the convergence of $\widetilde{\mathbb{P}}_{\boldsymbol{n}}$, $\widetilde{\mathbb{G}}_{\boldsymbol{n}}$ and $\widetilde{\mathbb{G}}_{\boldsymbol{n}}^*$ as $\underline{n}$ tends to infinity. More precisely, as with multisample U-statistics [see, e.g. 136, Section 12.2], we assume that there is an index $m \in \mathbb{N}^+$, left implicit hereafter, and increasing functions $g_1, ..., g_k$ such that for all j, $n_j = g_j(m) \to \infty$ as $m \to \infty$ (we also assume without loss of generality that for all $m \in \mathbb{N}^+$, $g_j(m+1) > g_j(m)$ for some j). The following theorem extends Theorems 4.1 and 4.2 to this set-up.

Theorem 4.6. *Suppose that Assumptions 4.2 and 4.5 hold and that for every $j = 1, ..., k$, there exists $\lambda_j \geq 0$ such that $\underline{n}/n_j \to \lambda_j \geq 0$. Then:*

1. *If Assumption 4.3 holds with F also satisfying $\widetilde{P}F < +\infty$, $\sup_{f \in \mathcal{F}} \left| \widetilde{\mathbb{P}}_{\boldsymbol{n}} f - \widetilde{P} f \right|$ tends to 0 almost surely and in L^1.*
2. *If $\mathbb{E}(N_{\mathbf{1}}^2) < +\infty$ and Assumption 4.4 holds with F also satisfying $\mathbb{E}\left(N_{\mathbf{1}} \sum_{\ell=1}^{N_{\mathbf{1}}} F^2(Y_{\mathbf{1},\ell}) \right) < +\infty$, the process $\widetilde{\mathbb{G}}_n$ converges weakly to a centered Gaussian process $\widetilde{\mathbb{G}}_\lambda$ on $\mathcal{F}$ as n tends to infinity. Moreover, the covariance kernel $\widetilde{K}$ of $\widetilde{\mathbb{G}}_\lambda$ satisfies:*

$$\widetilde{K}(f_1, f_2) = \sum_{j=1}^k \lambda_j \mathbb{C}ov \left(\sum_{\ell=1}^{N_{\mathbf{1}}} f_1(Y_{\mathbf{1},\ell}), \sum_{\ell=1}^{N_{\mathbf{2}_j}} f_2(Y_{\mathbf{2}_j,\ell}) \right), \tag{4.5}$$

   where $\mathbf{2}_j$ is the k-tuple with 2 in each entry but 1 in entry j.
3. *Under the same condition as in 2., the process $\widetilde{\mathbb{G}}_n^*$ converges weakly to $\widetilde{\mathbb{G}}_\lambda$, conditional on $(Y_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{N}^{+k}}$ and almost surely.*

Theorem 4.6 includes the case where $\lambda_j = 0$ for some j, corresponding to "strongly unbalanced" designs with different rates of convergence to $+\infty$ along the different dimensions of the array. In that case, only the dimensions with the slowest rate of convergence contribute to the asymptotic distribution, as can be seen in (4.5).

Because the $(n_j)_{j=1...k}$ are not all equal in general, Theorem 4.6 does not follow directly from Theorem 4.1, even if Assumption 4.5 is stronger than Assumption 4.1. We prove the result by showing a simpler and convenient version of the symmetrization lemma in this setting. We refer to Lemma S2 in Appendix B for more details.

4.4 Simulations and real data example

4.4.1 Monte Carlo simulations

We investigate in this section the finite sample properties of the bootstrap scheme considered above, by studying the coverage rate of confidence intervals based on this bootstrap. We consider dyadic data satisfying Assumption 4.1, with $N_{\boldsymbol{i}} = 1$ for all $\boldsymbol{i} \in \mathbb{I}_2$, and the following dependence structure:

$$Y_{i_1,i_2} = 1 + \mu(\varepsilon_{1i_1} + \varepsilon_{2i_2}) + \sqrt{0.5 - \mu^2} \left(\nu \varepsilon_{i_1,i_2}^S + \sqrt{2 - \nu^2} \varepsilon_{i_1,i_2} \right),$$

where the $(\varepsilon_{1i_1}, \varepsilon_{2i_1})_{i_1 \in \mathbb{N}^+}$, $(\varepsilon^S_{i_1,i_2})_{(i_1,i_2)\in \mathbb{I}_2}$ and $(\varepsilon_{i_1,i_2})_{(i_1,i_2)\in \mathbb{I}_2}$ are mutually independent and all standard normal variables. We impose Corr$(\varepsilon_{1i_1}, \varepsilon_{2i_1}) = 0.8$ and $\varepsilon^S_{i_1,i_2} = \varepsilon^S_{i_2,i_1}$. The parameter $\mu \in [0, 1/\sqrt{2}]$ represents the importance of individual versus pair factors, whereas $\nu \in [0, \sqrt{2}]$ represents the importance of symmetric versus asymmetric shocks. In the baseline scenario, we let $(\mu, \nu) = (\sqrt{0.2}, 1)$. We also consider two other scenarios, where respectively $(\mu, \nu) = (\sqrt{0.05}, 1)$ and $(\mu, \nu) = (\sqrt{0.2}, 0)$. Our parameter of interest θ_0 is the median of $Y_{1,2}$, which is thus equal to 1. Hereafter, we study inference on θ_0 based on the empirical median $\widehat{\theta}$, for $n \in \{10, 20, 40, 80\}$.

We inspect the performance of two different confidence intervals. The first is the symmetric interval $[\widehat{\theta} \pm q_{0.95}(|\widehat{\theta}^* - \widehat{\theta}|)]$, where $\widehat{\theta}^*$ denotes the bootstrap counterpart of $\widehat{\theta}$ and $q_\alpha(U)$ denotes the quantile of order α of U, conditional on the data $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in \mathbb{I}_{n,k}}$. The second is the percentile bootstrap interval $[q_{0.025}(\widehat{\theta}^*), q_{0.975}(\widehat{\theta}^*)]$. Given Theorem 4.4, both intervals are asymptotically valid.

Our results are displayed in Table 4.1. Our two confidence intervals have very good properties, overall, even for very small sample sizes. They appear to be slightly conservative for small n, but their coverage rate is already close to the nominal level for $n = 80$. The two confidence intervals are also very close to each other.

Scenario	n	Symmetric bootstrap CI	Percentile bootstrap CI
Baseline:	10	0.984	0.986
$\mu = \sqrt{0.2}$	20	0.977	0.979
$\nu = 1$	40	0.969	0.971
	80	0.961	0.961
baseline	10	0.984	0.986
but $\mu = \sqrt{0.05}$	20	0.977	0.979
	40	0.969	0.971
	80	0.961	0.961
baseline	10	0.98	0.983
but $\nu = 0$	20	0.971	0.972
	40	0.965	0.968
	80	0.962	0.961

Notes: 5,000 simulations, 200 bootstrap samples for each.

Table 4.1 – Coverage rates on the true median (nominal coverage: 95%)

4.4.2 Application to international trade data

Finally, we illustrate in this section the importance of accounting for dependence in real dyadic data. We revisit for that purpose [126], who estimate the so-called gravity equation for international trade. This gravity equation states that the trade volume T_{i_1,i_2} from country i_1 to country i_2 satisfies

$$T_{i_1,i_2} = \exp(\alpha_0) G_{i_1}^{\alpha_1} G_{i_2}^{\alpha_2} D_{i_1,i_2}^{\alpha_3} \exp\left(A_{i_1,i_2}\beta\right) \eta_{i_1,i_2}, \tag{4.6}$$

where G_i denotes country i's GDP, which would correspond to the mass of i in a traditional gravity equation, D_{i_1,i_2} denotes the distance between i_1 and i_2, A_{i_1,i_2} are additional control variables and η_{i_1,i_2} is an unobserved term.

We wish to estimate $\theta_0 = (\alpha_0, ..., \alpha_3, \beta')'$. The usual way to do so is to take the log in (4.6) and use the

OLS estimator. An issue, however, is that many trade volumes are equal to zero. This is the case for instance in 47.6% of the data used by [126]. Thus, one would either have to discard the corresponding data, resulting in a sample selection, or take an ad hoc transform such as $\log(\eta + x)$ for some $\eta > 0$ instead of the log. In both cases, the corresponding OLS estimator is no longer consistent.

Instead, [126] suggest to use the Poisson pseudo maximum likelihood (PPML for short) estimator $\widehat{\theta}$. The idea, formalized in [82], is that with i.i.d data, the PPML estimator is consistent and asymptotically normal for θ_0 even if $T_{\boldsymbol{i}}$ does not follow a Poisson model, provided that $\mathbb{E}[\eta_{\boldsymbol{i}}|X_{\boldsymbol{i}}] = 1$, with $X_{\boldsymbol{i}} = (1, \ln(G_{i_1}), \ln(G_{i_2}), \ln(D_{\boldsymbol{i}}), A_{\boldsymbol{i}})$. This is because the PPML estimator is based on the empirical counterpart of

$$\mathbb{E}\left[X_{\boldsymbol{i}}'\left(T_{\boldsymbol{i}} - \exp(X_{\boldsymbol{i}}\theta_0)\right)\right] = 0, \tag{4.7}$$

and this equality holds true if $\mathbb{E}[\eta_{\boldsymbol{i}}|X_{\boldsymbol{i}}] = 1$. Now, assuming as in [126] that the variables $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_2}$ are i.i.d. (with $Y_i = (T_{\boldsymbol{i}}, X_{\boldsymbol{i}})$) is restrictive. We suppose instead that Assumption 4.1 holds. Then Theorem 4.3 applies to this setting, implying that $\widehat{\theta}$ is still consistent and asymptotically normal in this case.[5] The rate of convergence and asymptotic variance are nonetheless different in the two cases, resulting in a different inference on θ_0.[6]

We use the same dataset as [126], which covers 136 countries for year 1990, and consider the exact same specification as the one they use in their Table 3. In this specification, the additional control variables $A_{\boldsymbol{i}}$ include exporter- and importer-level variables, namely their GDP per capita, a dummy variable equal to one if countries are landlocked and a remoteness index, which is the log of GDP-weighted average distance to all other countries. It also includes variables at the pair level, namely dummy variables for contiguity, common language, colonial tie, free-trade agreement and openness. This openness dummy is equal to one if at least one country is part of a preferential trade agreement. We refer to [126] for additional details.

Table 4.2 below presents the results. The first column displays the point estimates, which, as expected, are identical to those in [126]. The other columns display the p-values for the null hypothesis that θ_{0j}, the j-th component of θ_0, is equal to 0. In Column 2, these p-values are obtained assuming that the $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_2}$ are i.i.d. As in [126], the p-values are computed using asymptotic normality and estimators of the asymptotic variance. In Column 6, we report the p-values based on our bootstrap, hence supposing that Assumption 4.1 holds. We compute the p-value p_j for $\theta_{0j} = 0$ using $p_j = \mathbb{P}\left(|\widehat{\theta}_j^* - \widehat{\theta}_j| > |\widehat{\theta}_j| \big| (Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\right)$.

We consider in the other columns alternative forms of dependence that have been considered in applied work on similar data. Column 3 corresponds to pairwise clustering, where Y_{i_1,i_2} and Y_{i_2,i_1} may be dependent, but $Y_{\boldsymbol{i}}$ and $Y_{\boldsymbol{j}}$ are independent if $\boldsymbol{j}$ is not a permutation of $\boldsymbol{i}$. Column 4 corresponds to one-way clustering according to i_1, whereas Column 5 corresponds to one-way clustering according to i_2. In the former case, Y_{i_1,i_2} and Y_{i_1,i_3} may be dependent, but Y_{i_1,i_2} and Y_{i_1',i_3} are independent as soon as $i_1 \neq i_1'$. In the latter case, Y_{i_1,i_2} and Y_{i_3,i_2} may be dependent, but Y_{i_1,i_2} and $Y_{i_3,i_2'}$ are independent as soon as $i_2 \neq i_2'$. In Columns 3 to 5, we follow the usual practice of computing the p-values using the asymptotic normality of $\widehat{\theta}_j$ and estimators of the asymptotic variance under these various dependence structures.

[5] In this case, $\mathcal{H} = \{1, ..., \dim(X_{\boldsymbol{i}})\}$ and $\psi_{\theta,h}(Y_{\boldsymbol{i}}) = X_{h,\boldsymbol{i}}(T_{\boldsymbol{i}} - \exp(X_{\boldsymbol{i}}\theta_0))$. Then the key conditions 2 and 3 in Theorem 4.3 are satisfied as soon as Θ is bounded, see e.g. Example 19.7 in [136].

[6] The same application has been considered by [83], who shows, assuming convergence of a certain sample average, the asymptotic normality of the PPML estimator under the same dependence structure as ours. On the other hand, he neither considers bootstrap-based inference nor proves the consistency of his (asymptotic) variance estimator.

		p-values under different assumptions				
Variable	Estimator	i.i.d	PW	cl. E	cl. I	dyadic
Log(E's GDP)	0.732	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
Log(I's GDP)	0.741	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
Log(E's PCGDP)	0.157	0.003	$< 10^{-3}$	0.04	0.001	0.078
Log(I's PCGDP)	0.135	0.003	$< 10^{-3}$	0.004	0.055	0.076
Log of distance	-0.784	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
Contiguity	0.193	0.064	0.16	0.112	0.077	0.461
Common-language	0.746	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	0.056
Colonial-tie	0.025	0.867	0.902	0.891	0.882	0.952
Landlocked E	-0.863	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	0.004
Landlocked I	-0.696	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	0.011
E's remoteness	0.66	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	0.036
I's remoteness	0.562	$< 10^{-3}$	$< 10^{-3}$	0.003	0.004	0.105
P-T agreement	0.181	0.041	0.117	0.054	0.122	0.456
Openness	-0.107	0.416	0.522	0.498	0.453	0.771

Notes: data from [126], same specification as in their Table 3. "E", "I", "PCGDP", "P-T", "PW" and "cl." stand for exporter, importer, per capita GPD, preferential-trade, pairwise and clustering, respectively. The p-values for the last column were obtained with 1,000 bootstrap samples.

Table 4.2 – Point estimates of θ_0 and p-values of $\theta_{0j} = 0$ under different dependence assumptions

Using our bootstrap leads to much larger p-values than under the i.i.d. assumption. Only the log of GDP of the exporter and the importer and distance appear to be significant at the 10^{-3} levels, whereas five additional control variables are significant at that level under the i.i.d. assumption. In particular, common language and importer's remoteness are not even significant at the usual 5% level. Interestingly, there is also a gap between assuming one-way clustering, either at the exporter or at the importer level, and assuming to have a jointly exchangeable and dissociated array. In the former case, we still have seven variables that are significant at the 10^{-3} levels. Confidence intervals, not displayed here, lead to similar conclusions. In particular, compared to the average length of i.i.d.-based 95% confidence intervals, those based on pairwise clustering are only 8% wider. Those based on one-way clustering on exporters (resp. importers) are 20% (resp. 17%) larger. On the other hand, those based on Assumption 4.1 are 136% wider.

Finally, note that [126] also consider a model with country fixed effects. In such a case, and even if the data are i.i.d., the PPML estimator has a non-negligible bias compared to its standard error [94], thus leading to distorted inference if not accounted for. [94] considers an alternative estimator and shows that it is asymptotically normal and unbiased if the error terms are i.i.d. Theorem 4.3 above does not apply directly to this non-standard estimator, but we conjecture that it is still asymptotically normal, and the bootstrap valid, if the data are jointly exchangeable and dissociated.

4.5 Conclusion

While polyadic data are increasingly used in applied work, and empirical researchers routinely account for multiway clustering when computing standard errors, the statistical theory behind these forms of dependence has lagged behind. We first contribute to this literature by linking these dependence

structures to jointly and separately exchangeable arrays. Using representation results for such arrays, we then prove uniform laws of large numbers and central limit theorems. These results imply consistency and asymptotic normality of various nonlinear estimators under such dependence. We also establish the general validity of natural extensions of the standard nonparametric bootstrap to such arrays. Our application shows that using those bootstrap schemes may make a large difference compared to assuming i.i.d. data or clustering along a single dimension, as has often been done.

One caveat is that for the bootstrap confidence intervals to be valid, the asymptotic variance of the estimator should be positive. This may not be the case, for instance if the data $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$ are actually i.i.d. Inference based on the wild bootstrap without this positivity condition has been studied for sample averages under multiway clustering by [109]. How to conduct inference for jointly exchangeable arrays or nonlinear estimators under multiway clustering without this positivity condition remains an avenue for future research.

4.6 Appendix A

4.6.1 Proof of Lemma 2.2

The general idea of the proof of (4.3) in the i.i.d. case is first to bound the initial expectation by another one involving a sum of independent differences of identically distributed variables. By symmetry, these differences can be multiplied by Rademacher variables without affecting the expectation. We follow this general strategy here, but complications arise because of dependence in the $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_{n,k}}$.

Specifically, we proceed in four steps. In the first step, we obtain an upper bound with a sum of differences that are identically distributed but not independent. Roughly speaking, they are nonetheless "less dependent", as we "decouple" the random variables appearing in the AHK representation (4.1) by introducing independent copies of them (see inequality (4.11) below). In the second step, using a telescopic sum, we further bound our expectation of interest by another one involving sums of differences that are independent, conditional on a suitable σ-algebra. The third step is the symmetrisation step itself, where Rademacher variables are introduced. The fourth step concludes by combining the previous steps. Note that the key decoupling inequality (4.11) is given separately in Lemma 4.3, as it may be of independent interest.

First step: decoupling

For any $j \in \mathbb{N}$, let $\left(U_A^{(j)}\right)_{A\subset\mathbb{N}^+:1\leq|A|\leq r}$ and $\left(V_A^{(j)}\right)_{A\subset\mathbb{N}^+:1\leq|A|\leq r}$ denote some independent copies of the $(U_A)_{A\subset\mathbb{N}^+:1\leq|A|\leq r}$. Let $Y_{\boldsymbol{i}}^{(k)} = \tau\left(\left(U^{(0)}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\cup_{j=1}^k\mathcal{E}_j}\right)$ and

$$Y_{\boldsymbol{i}}^{(r)} = \tau\left(\left(U^{(0)}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\cup_{j=1}^r\mathcal{E}_j}, \left(V^{(0)}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\cup_{j=r+1}^k\mathcal{E}_j}\right).$$

Because $\mathbb{E}\left[f\left(Y_{\mathbf{1}}\right)\right] = \mathbb{E}\left[f\left(Y_{\boldsymbol{i}}^{(k)}\right)|Y_{\boldsymbol{i}}^{(0)}\right]$ and $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k} \overset{d}{=} \left(Y_{\boldsymbol{i}}^{(0)}\right)_{\boldsymbol{i}\in\mathbb{I}_k}$, we obtain, by Jensen's inequality and Lemma S1,

$$\begin{aligned}
&\mathbb{E}\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} f\left(Y_{\boldsymbol{i}}\right) - \mathbb{E}\left[f\left(Y_{\mathbf{1}}\right)\right]\right|\right)\right] \\
\leq&\mathbb{E}\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} f\left(Y_{\boldsymbol{i}}^{(0)}\right) - f\left(Y_{\boldsymbol{i}}^{(k)}\right)\right|\right)\right] \\
\leq&\frac{1}{k}\sum_{r=1}^k\mathbb{E}\left[\Phi\left(k\sup_{f\in\mathcal{F}}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\in\mathbb{I}_{n,k}} f\left(Y_{\boldsymbol{i}}^{(r-1)}\right) - f\left(Y_{\boldsymbol{i}}^{(r)}\right)\right|\right)\right].
\end{aligned} \tag{4.8}$$

For $\boldsymbol{i}\in\mathbb{R}^k$ and $\pi\in\mathfrak{S}_k$, let $\boldsymbol{i}_\pi = (i_{\pi(1)},...,i_{\pi(k)})$. For any $f\in\mathcal{F}$, let also

$$\overline{f}\left(\left(U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\cup_{r=1}^k\mathcal{E}_r}\right) = \frac{1}{k!}\sum_{\pi\in\mathfrak{S}_k} f(Y_{\boldsymbol{i}_\pi}).$$

Note that $\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\overline{f}\left(\left(U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\cup_{r=1}^k\mathcal{E}_r}\right) = \sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} f(Y_{\boldsymbol{i}})$ and if the components of $\boldsymbol{i}'$ are a permutation of those of $\boldsymbol{i}$ we have

$$\overline{f}\left(\left(U_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\cup_{r=1}^k\mathcal{E}_r}\right) = \overline{f}\left(\left(U_{\{\boldsymbol{i}'\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\cup_{r=1}^k\mathcal{E}_r}\right). \tag{4.9}$$

For $r = 1,...,k$, let $\overline{\mathcal{E}}_r = \cup_{j=r+1}^k\mathcal{E}_j$ and $\underline{\mathcal{E}}_r = \cup_{j=1}^{r-1}\mathcal{E}_r$. Let $\mathcal{U}^r$ be the σ-algebra generated by the variables $(U^{(0)}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{(\boldsymbol{i},\boldsymbol{e})\in\mathbb{I}_{n,k}\times\underline{\mathcal{E}}_r}$ and $(V^{(0)}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+})_{(\boldsymbol{i},\boldsymbol{e})\in\mathbb{I}_{n,k}\times\overline{\mathcal{E}}_r}$. For any $j\in\mathbb{N}$, $\boldsymbol{i}\in\mathbb{I}_{n,k}$ and $\boldsymbol{e}\in\cup_{j'=1}^k\mathcal{E}_{j'}$, let $W^{(j)}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+} = \left(U^{(j)}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}, V^{(j)}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)$.

As we will reason conditional on $\mathcal{U}^r$, let us use $\overline{f}_{r,\boldsymbol{i}}(\boldsymbol{w})$ as a shortcut for

$$\overline{f}\left(\left(U^{(0)}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\underline{\mathcal{E}}_r}, \boldsymbol{w}, \left(V^{(0)}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\overline{\mathcal{E}}_r}\right),$$

for any vector $\boldsymbol{w}\in\mathbb{R}^{|\mathcal{E}_r|}$. Let us also define

$$\begin{aligned}&\Delta\overline{f}_{r,\boldsymbol{i}}\left(\left(W^{(0)}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r}\right)\\ =&k\frac{(n-k)!}{n!}\left[\overline{f}_{r,\boldsymbol{i}}\left(\left(U^{(0)}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r}\right)-\overline{f}_{r,\boldsymbol{i}}\left(\left(V^{(0)}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r}\right)\right].\end{aligned}$$

Then, by definition of $Y^{(r)}_{\boldsymbol{i}}$ and $\Delta\overline{f}_{r,\boldsymbol{i}}$,

$$\begin{aligned}&\mathbb{E}\left[\Phi\left(k\sup_{f\in\mathcal{F}}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}f\left(Y^{(r-1)}_{\boldsymbol{i}}\right)-f\left(Y^{(r)}_{\boldsymbol{i}}\right)\right|\right)\middle|\mathcal{U}^r\right]\\ =&\mathbb{E}\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\Delta\overline{f}_{r,\boldsymbol{i}}\left(\left(W^{(0)}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r}\right)\right|\right)\middle|\mathcal{U}^r\right].\end{aligned}\tag{4.10}$$

Remark that the first result in Lemma 4.3 applies conditional on $\mathcal{U}^r$. Then, letting $K_{1,r}=\left(3|\mathcal{E}_r|^{|\mathcal{E}_r|}\right)^{|\mathcal{E}_r|-1}$ and ℓ be an arbitrary bijection from $\mathcal{E}_r$ to $\{1,...,|\mathcal{E}_r|\}$, we obtain

$$\begin{aligned}&\mathbb{E}\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\Delta\overline{f}_{r,\boldsymbol{i}}\left(\left(W^{(0)}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r}\right)\right|\right)\middle|\mathcal{U}^r\right]\\ \leq&\mathbb{E}\left[\Phi\left(K_{1,r}\sup_{f\in\mathcal{F}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\Delta\overline{f}_{r,\boldsymbol{i}}\left(\left(W^{(\ell(\boldsymbol{e}))}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r}\right)\right|\right)\middle|\mathcal{U}^r\right].\end{aligned}\tag{4.11}$$

Second step: telescoping sum

Let $\prec$ be a total order on $\mathcal{E}_r$. We note $\boldsymbol{e}\preceq\boldsymbol{e}'$ if $\boldsymbol{e}\prec\boldsymbol{e}'$ or $\boldsymbol{e}=\boldsymbol{e}'$. For every $(\boldsymbol{e},\boldsymbol{e}')\in\mathcal{E}_r^2$ let

$$\overline{W}^{(\ell,\boldsymbol{e})}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}=\left|\begin{array}{ll}\left(U^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+},U^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\right) & \text{if } \boldsymbol{e}'\prec\boldsymbol{e}\\ \left(V^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+},V^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\right) & \text{if } \boldsymbol{e}'\succ\boldsymbol{e}\\ \left(U^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+},V^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\right) & \text{if } \boldsymbol{e}'=\boldsymbol{e}.\end{array}\right.$$

Then, for any $\boldsymbol{e}\in\mathcal{E}_r$,

$$\begin{aligned}\left(\overline{W}^{(\ell,\boldsymbol{e})}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}=&\left(U^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\mathbb{1}_{\{\boldsymbol{e}'\preceq\boldsymbol{e}\}}+V^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\mathbb{1}_{\{\boldsymbol{e}'\succ\boldsymbol{e}\}},\right.\\ &\left.U^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\mathbb{1}_{\{\boldsymbol{e}'\prec\boldsymbol{e}\}}+V^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\mathbb{1}_{\{\boldsymbol{e}'\succeq\boldsymbol{e}\}}\right)_{\boldsymbol{e}'\in\mathcal{E}_r},\end{aligned}\tag{4.12}$$

and $\left(\overline{W}^{(\ell,\boldsymbol{e}')}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}=\left(W^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}$. $\Delta\overline{f}_{r,\boldsymbol{i}}\left(\left(W^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}\right)$ can be decomposed into the following telescoping sum:

$$\Delta\overline{f}_{r,\boldsymbol{i}}\left(\left(W^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}\right)=\sum_{\boldsymbol{e}\in\mathcal{E}_r}\Delta\overline{f}_{r,\boldsymbol{i}}\left(\left(\overline{W}^{(\ell,\boldsymbol{e})}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}\right).$$

By Lemma S1, we obtain, with $K_{2,r}=|\mathcal{E}_r|K_{1,r}$,

$$\begin{aligned}&\mathbb{E}\left[\Phi\left(K_{1,r}\sup_{f\in\mathcal{F}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\Delta\overline{f}_{r,\boldsymbol{i}}\left(\left(W^{\ell(\boldsymbol{e}')}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}\right)\right|\right)\middle|\mathcal{U}^r\right]\\ \leq&\frac{1}{|\mathcal{E}_r|}\sum_{\boldsymbol{e}\in\mathcal{E}_r}\mathbb{E}\left[\Phi\left(K_{2,r}\sup_{f\in\mathcal{F}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\Delta\overline{f}_{r,\boldsymbol{i}}\left(\left(\overline{W}^{(\ell,\boldsymbol{e})}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}\right)\right|\right)\middle|\mathcal{U}^r\right].\end{aligned}\tag{4.13}$$

Third step: symmetrization

For any $\boldsymbol{e} \in \mathcal{E}_r$, let $\mathcal{U}^r_{\ell,\boldsymbol{e}}$ be the σ-algebra generated by the same variables as $\mathcal{U}^r$, $(U^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+})_{(\boldsymbol{i}\times\boldsymbol{e}')\in\mathbb{I}_{n,k}\times\underline{\mathcal{E}}_r:\boldsymbol{e}'\prec\boldsymbol{e}}$ and $(V^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+})_{(\boldsymbol{i},\boldsymbol{e}')\in\mathbb{I}_{n,k}\times\underline{\mathcal{E}}_r:\boldsymbol{e}'\succ\boldsymbol{e}}$. Let $\overrightarrow{\mathbb{I}_{n,k}} = \{(i_1,i_2,...,i_k)\in\{1,...,n\}^k : i_1 < i_2 < ... < i_k\} \subset \mathbb{I}_{n,k}$ and $\mathfrak{S}_k$ be the set of permutations of $\{1,...,k\}$. For any $\boldsymbol{i} = (i_1,...,i_k)\in\mathbb{N}^k$ and $\pi\in\mathfrak{S}_k$, let $\boldsymbol{i}_\pi$ denote $(i_{\pi(1)},...,i_{\pi(k)})$. For any $\boldsymbol{i}\in\mathbb{I}_r$ and $\boldsymbol{e}\in\mathcal{E}_r$, let $\boldsymbol{i}^{\boldsymbol{e}}$ be the k-dimensional vector with component i_1 in the first non-null entry of $\boldsymbol{e}$, i_2 in the second non-null entry of $\boldsymbol{e}$ and so on. Similarly, for any $\boldsymbol{i}\in\mathbb{I}_{k-r}$ and $\boldsymbol{e}\in\mathcal{E}_r$, let $\boldsymbol{i}^{(\mathbf{1}-\boldsymbol{e})}$ be the k-dimensional vector with component i_1 at the first null entry of $\boldsymbol{e}$, i_2 at the second null entry of $\boldsymbol{e}$ and so on. For instance, if $k=5$, $r=3$, $\boldsymbol{i}=(6,9,2)$, $\boldsymbol{i}'=(7,3)$ and $\boldsymbol{e}=(0,1,1,0,1)$, we obtain $\boldsymbol{i}^{\boldsymbol{e}}=(0,6,9,0,2)$ and $\boldsymbol{i}'^{(\mathbf{1}-\boldsymbol{e})}=(7,0,0,3,0)$.

For every $\boldsymbol{e}\in\mathcal{E}_r$, we have

$$\mathbb{I}_{n,k} = \left\{\boldsymbol{i}^{\boldsymbol{e}}_\pi + \boldsymbol{i}'^{(\mathbf{1}-\boldsymbol{e})} : \boldsymbol{i}\in\overrightarrow{\mathbb{I}_{n,r}}, \pi\in\mathfrak{S}_r, \boldsymbol{i}'\in\overrightarrow{(\{1,...,n\}\backslash\{\boldsymbol{i}\})^{k-r}}\right\}. \tag{4.14}$$

Thus,

$$\begin{aligned}
&\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} \Delta\overline{f}_{r,\boldsymbol{i}}\left(\left(\overline{W}^{(\ell,\boldsymbol{e})}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}\right)\\
&= \sum_{\boldsymbol{i}\in\overrightarrow{\mathbb{I}_{n,r}}}\sum_{\boldsymbol{i}'\in\overrightarrow{(\{1,...,n\}\backslash\{\boldsymbol{i}\})^{k-r}}}\sum_{\pi\in\mathfrak{S}_r}\Delta\overline{f}_{r,\boldsymbol{i}^{\boldsymbol{e}}_\pi+\boldsymbol{i}'^{(\mathbf{1}-\boldsymbol{e})}}\left(\left(\overline{W}^{(\ell,\boldsymbol{e})}_{\{(\boldsymbol{i}^{\boldsymbol{e}}_\pi+\boldsymbol{i}'^{(\mathbf{1}-\boldsymbol{e})})\odot\boldsymbol{e}'\}^+}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}\right).
\end{aligned}$$

With this new indexation of the sum on $\boldsymbol{i}$ and reasoning conditional on $\mathcal{U}^r_{\ell,\boldsymbol{e}}$, the triple sum above can be rewritten as a sum of $n!/[(n-r)!r!]$ symmetric and independent terms. Hence, it is equal in distribution to

$$\sum_{\boldsymbol{i}\in\overrightarrow{\mathbb{I}_{n,r}}}\varepsilon_{\{\boldsymbol{i}\}}\sum_{\boldsymbol{i}'\in\overrightarrow{(\{1,...,n\}\backslash\{\boldsymbol{i}\})^{k-r}}}\sum_{\pi\in\mathfrak{S}_r}\Delta\overline{f}_{r,\boldsymbol{i}^{\boldsymbol{e}}_\pi+\boldsymbol{i}'^{(\mathbf{1}-\boldsymbol{e})}}\left(\left(\overline{W}^{(\ell,\boldsymbol{e})}_{\{(\boldsymbol{i}^{\boldsymbol{e}}_\pi+\boldsymbol{i}'^{(\mathbf{1}-\boldsymbol{e})})\odot\boldsymbol{e}'\}^+}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}\right),$$

where the $(\varepsilon_A)_{A\subset\{1,...,n\}}$ are i.i.d. Rademacher variables. For every $\boldsymbol{i}\in\overrightarrow{\mathbb{I}_{n,r}}$ and any $\pi\in\mathfrak{S}_r$, we have $\{\boldsymbol{i}\} = \{(\boldsymbol{i}^{\boldsymbol{e}}_\pi + \boldsymbol{i}'^{(\mathbf{1}-\boldsymbol{e})})\odot\boldsymbol{e}\}^+$. Hence, using (4.14) again,

$$\begin{aligned}
&\sum_{\boldsymbol{i}\in\overrightarrow{\mathbb{I}_{n,r}}}\varepsilon_{\{\boldsymbol{i}\}}\sum_{\boldsymbol{i}'\in\overrightarrow{(\{1,...,n\}\backslash\{\boldsymbol{i}\})^{k-r}}}\sum_{\pi\in\mathfrak{S}_r}\Delta\overline{f}_{r,\boldsymbol{i}^{\boldsymbol{e}}_\pi+\boldsymbol{i}'^{(\mathbf{1}-\boldsymbol{e})}}\left(\left(\overline{W}^{(\ell,\boldsymbol{e})}_{\{(\boldsymbol{i}^{\boldsymbol{e}}_\pi+\boldsymbol{i}'^{(\mathbf{1}-\boldsymbol{e})})\odot\boldsymbol{e}'\}^+}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}\right)\\
&= \sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\Delta\overline{f}_{r,\boldsymbol{i}}\left(\left(\overline{W}^{(\ell,\boldsymbol{e})}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}\right).
\end{aligned}$$

Furthermore, for every $\boldsymbol{e}\in\mathcal{E}_r$, by (4.12),

$$\begin{aligned}
&\frac{n!}{k(n-k)!}\Delta\overline{f}_{r,\boldsymbol{i}}\left(\left(\overline{W}^{(\ell,\boldsymbol{e})}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}\right)\\
=&\overline{f}_{r,\boldsymbol{i}}\left(\left(U^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\mathbb{1}_{\{\boldsymbol{e}'\preceq\boldsymbol{e}\}} + V^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\mathbb{1}_{\{\boldsymbol{e}'\succ\boldsymbol{e}\}}\right)_{\boldsymbol{e}\in\mathcal{E}_r}\right)\\
&-\overline{f}_{r,\boldsymbol{i}}\left(\left(U^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\mathbb{1}_{\{\boldsymbol{e}'\prec\boldsymbol{e}\}} + V^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\mathbb{1}_{\{\boldsymbol{e}'\succeq\boldsymbol{e}\}}\right)_{\boldsymbol{e}\in\mathcal{E}_r}\right).
\end{aligned}$$

Since for every $(j,j')\in\mathbb{N}^2$, $(U^{(j)}_A)_{A\subseteq\{1,...,n\}}$ and $(V^{(j')}_A)_{A\subseteq\{1,...,n\}}$ are equal in distribution and independent and $(U^{(j)}_A)_{A\subseteq\{1,...,n\}} \perp\!\!\!\perp (U^{(j')}_A)_{A\subseteq\{1,...,n\}}$ whenever $j\neq j'$, we obtain, conditional on $\mathcal{U}^r$,

$$\begin{aligned}
&\left(\left(U^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\mathbb{1}_{\{\boldsymbol{e}'\preceq\boldsymbol{e}\}} + V^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\mathbb{1}_{\{\boldsymbol{e}'\succ\boldsymbol{e}\}}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}\right)_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\\
&\overset{d}{=}\left(\left(U^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\mathbb{1}_{\{\boldsymbol{e}'\prec\boldsymbol{e}\}} + V^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\mathbb{1}_{\{\boldsymbol{e}'\succeq\boldsymbol{e}\}}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}\right)_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\\
&\overset{d}{=}\left(\left(U^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}\right)_{\boldsymbol{i}\in\mathbb{I}_{n,k}}.
\end{aligned}$$

Then, by independence between $(\varepsilon_A)_{A\subset\mathbb{N}^+:1\leq|A|\leq k}$ and $(U_A^{(j)},V_A^{(j)})_{j\in\mathbb{N},A\subset\mathbb{N}^+:1\leq|A|\leq k}$ and the triangle and Jensen inequalities

$$\begin{aligned}&\frac{1}{|\mathcal{E}_r|}\sum_{\boldsymbol{e}\in\mathcal{E}_r}\mathbb{E}\left[\Phi\left(K_{2,r}\sup_{f\in\mathcal{F}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\Delta\overline{f}_{r,\boldsymbol{i}}\left(\left(\overline{W}^{(\ell,\boldsymbol{e})}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}\right)\right|\right)\middle|\mathcal{U}^r\right]\\ \leq&\frac{1}{|\mathcal{E}_r|}\sum_{\boldsymbol{e}\in\mathcal{E}_r}\mathbb{E}\left[\Phi\left(K_{3,r}\sup_{f\in\mathcal{F}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\overline{f}_{r,\boldsymbol{i}}\left(\left(U^{(\ell(\boldsymbol{e}'))}_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}\right)_{\boldsymbol{e}'\in\mathcal{E}_r}\right)\right|\right)\middle|\mathcal{U}^r\right],\end{aligned}\tag{4.15}$$

where $K_{3,r}=2k\frac{(n-k)!}{n!}K_{2,r}$.

Fourth step: conclusion

Combining Equations (4.8), (4.10), (4.11), (4.13), (4.15) and using the expressions of $K_{1,r}$, $K_{2,r}$ and $K_{3,r}$, we finally obtain

$$\begin{aligned}&\mathbb{E}\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}f\left(Y_{\boldsymbol{i}}\right)-\mathbb{E}\left[f\left(Y_{(\boldsymbol{1})}\right)\right]\right|\right)\right]\\ \leq&\frac{1}{k}\sum_{r=1}^{k}\frac{1}{|\mathcal{E}_r|}\sum_{\boldsymbol{e}\in\mathcal{E}_r}\mathbb{E}\left[\Phi\left(C_{r,k}\frac{(n-k)!}{n!}\sup_{f\in\mathcal{F}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{\boldsymbol{i}\odot\boldsymbol{e}'\}^+}f\left(Y^r_{\boldsymbol{i}}\right)\right|\right)\right],\end{aligned}$$

with $C_{r,k}=2k|\mathcal{E}_r|\left(3|\mathcal{E}_r|^{\mathcal{E}_r}\right)^{|\mathcal{E}_r|-1}$ and

$$Y^r_{\boldsymbol{i}}=\tau\left(\left(U^{(0)}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\underline{\mathcal{E}}_r},\left(U^{(\ell(\boldsymbol{e}))}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r},\left(V^{(0)}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\overline{\mathcal{E}}_r}\right).$$

By construction of the $\left(U_A^{(j)}\right)_{A\subset\mathbb{N}^+:1\leq|A|\leq k}$ and $\left(V_A^{(0)}\right)_{A\subset\mathbb{N}^+:r+1\leq|A|\leq k}$, $(Y^r_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$ is jointly exchangeable and dissociated, with marginal distribution P. This concludes the proof.

4.6.1.1 A decoupling inequality

The proof of Lemma 4.2 crucially hinges upon the following decoupling inequality, which may be of independent interest. Hereafter, we let $\mathcal{A}_r=\{A\subseteq\{1,...n\}:|A|=r\}$.

Lemma 4.3. *Let $r\leq k$, $(W_A)_{A\in\mathcal{A}_r}$ be a family of i.i.d. random variables with values in a Polish space $\mathcal{S}$ and $\left(W_A^{(j)}\right)_{A\in\mathcal{A}_r}$, $j=1,...,|\mathcal{E}_r|$ be some independent copies of this family. Let Φ be a non-decreasing convex function from $\mathbb{R}^+$ to $\mathbb{R}$ and ℓ be a bijection from $\mathcal{E}_r$ to $\{1,...,|\mathcal{E}_r|\}$. Let $\mathcal{H}$ be a pointwise measurable class of functions from $\mathcal{S}^{|\mathcal{E}_r|}\times\mathbb{I}_{n,k}$ to $\mathbb{R}$ such that $\mathbb{E}\left(\sup_{h\in\mathcal{H}}\left|h\left(\left(W_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r},\boldsymbol{i}\right)\right|\right)<\infty$. Finally, let $L_r=\left(3|\mathcal{E}_r|^{|\mathcal{E}_r|}\right)^{|\mathcal{E}_r|-1}$. Then*

$$\begin{aligned}&\mathbb{E}\Phi\left(\sup_{h\in\mathcal{H}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}h\left(\left(W_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r},\boldsymbol{i}\right)\right|\right)\\ \leq&\mathbb{E}\Phi\left(L_r\sup_{h\in\mathcal{H}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}h\left(\left(W^{(\ell(\boldsymbol{e}))}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r},\boldsymbol{i}\right)\right|\right).\end{aligned}$$

The proof is given in Appendix B. This result generalizes the decoupling inequality for U-statistics of [55] to our setting. As with U-statistics, it is possible to obtain a reverse inequality if $r\in\{1,k-1,k\}$ and $\pi\mapsto h\left(\left(W_{\{\boldsymbol{i}_\pi\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r},\boldsymbol{i}_\pi\right)$ is constant on $\mathfrak{S}_k$, for all $h\in\mathcal{H}$. With such a reverse inequality, it is possible to replace $Y^r_{\boldsymbol{i}}$ by $Y_{\boldsymbol{i}}$ in Lemma 4.2. It is unclear to us, however, whether this reverse inequality still holds if $r\notin\{1,k-1,k\}$ (implying $k\geq4$). The key argument for the reverse inequality in [55] is that

by the symmetry condition above, we can replace $h\left(\left(W_{\{\boldsymbol{i}_\pi \odot \boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}_\pi\right)$ by an average over $k!$ terms. However, for the proof to extend to our setting, one would need an average over $|\mathcal{E}_r|!$ terms. This is not possible in general when $|\mathcal{E}_r| > k$, which is the case when $r \notin \{1, k-1, k\}$.

4.7 Appendix B

We prove in this appendix all the results presented in the paper, except the symmetrization lemma (Lemma 4.2). The first section gathers the proofs of the results in the jointly exchangeable case with one unit per cell (Section 4.2 of the paper), while the second section focuses on the proofs of the extensions. Section 3 collects all the technical lemmas.

To ease the reading, we first summarize the notation we use throughout the proofs. Objects introduced in a single proof are defined therein directly and not reported here. We recall that k denotes the dimension of the array of data. Also, bootstrap counterparts appear with a star.

Subsets or elements of $\mathbb{N}^k$

A^+	$A \cap (0, +\infty)$, for any $A \subset \mathbb{R}$.
$\overline{A}$	$\{\boldsymbol{i} \in A : i_j \neq i_{j'}$ if $j \neq j'\}$, for any $A \subset \mathbb{N}^{+k}$.
$\overrightarrow{A}$	$\{\boldsymbol{i} \in \overline{A} : i_j < i_{j'}$ if $j < j'\}$ for any $A \subset \mathbb{N}^{+k}$.
$\lvert A\rvert$	the cardinal of $A \subset \mathbb{N}^{+k}$.
$\mathfrak{S}(A)$	The set of permutations on A.
$\mathfrak{S}_r$	$\mathfrak{S}(\{1, ..., r\})$
$\mathbb{I}_k$	$\overline{\mathbb{N}^{+k}}$.
$\mathbb{I}_{n,k}$	$\overline{\{1, ..., n\}^k}$.
$\mathcal{E}_r$	$\{\boldsymbol{e} \in \{0; 1\}^k : \sum_{j=1}^k e_j = r\}$ for $r = 1, ..., k$.
$\boldsymbol{i}$	element of $\mathbb{I}_k$ or $\mathbb{N}^{+k}$, with component $(i_1, ..., i_k)$.
$\{\boldsymbol{i}\}$	the set of distinct elements of $\boldsymbol{i} = (i_1, ..., i_k) \in \mathbb{N}^k$.
$\boldsymbol{e}$	element of $\{0, 1\}^k$.
$\boldsymbol{i}^{\boldsymbol{e}}$	for $\boldsymbol{i} \in \mathbb{I}_{n,r}$ and $\boldsymbol{e} \in \mathcal{E}_r$, the k-dimensional vector with component i_1 at the first non-null entry of $\boldsymbol{e}$, i_2 at the second non-null entry of $\boldsymbol{e}$ and so on.[7]
$\mathbf{0}$	$(0, ..., 0)$
$\mathbf{1}$	$(1, ..., k)$ except in Section 4.7.2.3 and Lemmas S4.5, S4.7 and S4.9, where $\mathbf{1} = (1, ..., 1)$.
$\mathbf{2}_r$	element of $\mathbb{N}^k$ with 2 at each component but 1 at its rth component.
$\boldsymbol{i}_\pi$	$(i_{\pi(1)}, ..., i_{\pi(r)})$, for any $\boldsymbol{i} \in \mathbb{N}^r$ and $\pi \in \mathfrak{S}_r$.
$\odot$	the Hadamard product, i.e. $\boldsymbol{i} \odot \boldsymbol{e} = (i_1 e_1, ..., i_k e_k)$.

Sample and random variables

n	Number of units in the population.
$\boldsymbol{n}$	$(n_1, ...n_k)$, with n_j the number of clusters in the j-th dimension in Section 4.7.2.3.
$\Pi_{\boldsymbol{n}}$	$\prod_{j=1}^k n_j$.
$\widetilde{Y}_{\boldsymbol{i}}$	$(N_{\boldsymbol{i}}, (Y_{\boldsymbol{i},\ell})_{\ell=1...N_{\boldsymbol{i}}})$.
$(\varepsilon_A)_{A\in\mathcal{A}}$	Mutually independent Rademacher random variables (i.e., with values 1 or -1 with probability $1/2$), for any set $\mathcal{A}$.
$(Y^r_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$	jointly exchangeable array defined in Lemma 4.2 with marginal distribution P.

[7] For instance if $k = 5$, $r = 3$, $\boldsymbol{i} = (6, 9, 2)$ and $\boldsymbol{e} = (0, 1, 1, 0, 1)$, we obtain $\boldsymbol{i}^{\boldsymbol{e}} = (0, 6, 9, 0, 2)$.

$(\widetilde{Y}_{\boldsymbol{i}}^{r})_{\boldsymbol{i}\in\mathbb{I}_k}$	same as $(Y_{\boldsymbol{i}}^{r})_{\boldsymbol{i}\in\mathbb{I}_k}$, but when applying Lemma 4.2 to $\widetilde{\mathcal{F}}$ and $(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$ instead of $\mathcal{F}$ and $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$.
$(\widetilde{Y}_{\boldsymbol{i}}^{r,r'})_{\boldsymbol{i}\in\mathbb{I}_k}$	same as $(\widetilde{Y}_{\boldsymbol{i}}^{r})_{\boldsymbol{i}\in\mathbb{I}_k}$.

Functions and classes of functions

Id	The identity function.
$\mathcal{D}$	$\cup_{n\in\mathbb{N}}(\{n\}\times\mathcal{Y}^n)$.
$\widetilde{f}$	for any function f from $\mathcal{Y}$ to $\mathbb{R}$, the function from $\mathcal{D}$ to $\mathbb{R}$ defined by $\widetilde{f}(n,y_1,...,y_n)=\sum_{\ell=1}^{n} f(y_\ell)$.
$\widetilde{\mathcal{F}}$	$\left\{\widetilde{f}: f\in\mathcal{F}\right\}$.
$\mathcal{F}^2$	$\{f^2: f\in\mathcal{F}\}$, for any class of functions $\mathcal{F}$.
$\mathcal{F}\times\mathcal{G}$	$\{(f,g): f\in\mathcal{F}, g\in\mathcal{G}\}$.
$\mathcal{F}_\delta$	$\left\{h=f_1-f_2:(f_1,f_2)\in\mathcal{F}\times\mathcal{F}, \mathbb{E}\left[(f_1(Y_{\ell,\mathbf{1}})-f_2(Y_{\ell,\mathbf{1}}))^2\right]\leq\delta^2\right\}$.
$\mathcal{F}_\infty$	$\{h=f_1-f_2:(f_1,f_2)\in\mathcal{F}\times\mathcal{F}\}$.
$N(\eta,\mathcal{F},\|.\|)$	the minimal number of $\|.\|$-closed balls of radius η with centers in $\mathcal{F}$ needed to cover $\mathcal{F}$.
$J_{\mathcal{F}}(u)$	$\int_0^u \sup_Q \sqrt{\log N(\eta\|\|F\|\|_{Q,2},\mathcal{F},\|.\|_{Q,2})}d\eta$, where the supremum is taken over the set of probability measures with finite support.

Probability measures and norms

Note that we sometimes need to evaluate random variables at some specific value of the probability space. We denote by ω elements of this probability space Ω.

Qf	$\int f dQ$, for any probability measure Q.
P	the probability distribution of $Y_{\boldsymbol{i}}$.
$\mathbb{P}_n$, $\mathbb{P}'_n$	$\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\delta_{Y_{\boldsymbol{i}}}$ and $\frac{1}{n^k}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\delta_{Y_{\boldsymbol{i}}}$, respectively.
$\mathbb{P}_n^*$	$\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} W_{\boldsymbol{i}} f(Y_{\boldsymbol{i}})$, where $W_{\boldsymbol{i}}$ is the bootstrap weight of $\boldsymbol{i}$.
$\mathbb{P}_n^r$	$\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\delta_{Y_{\boldsymbol{i}}^r}$
$\|\|g\|\|_{\mu,r}$	$\left(\int\|g\|^r d\mu\right)^{1/r}$ for μ a measure and $r\geq 1$
$\|\|f\|\|_{\boldsymbol{e},M,1}$	$\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\overrightarrow{\mathbb{I}_{n,r}}}\left\|\sum_{\pi\in\mathfrak{S}_r}\sum_{\boldsymbol{i}'\in\overrightarrow{(\{1,...,n\}\setminus\{\boldsymbol{i}\})^{k-r}}} f\left(Y^r_{(\boldsymbol{i}_\pi)^{\boldsymbol{e}}+\boldsymbol{i}'(\mathbf{1}-\boldsymbol{e})}\right)\mathbb{1}_{\left\{F\left(Y^r_{(\boldsymbol{i}_\pi)^{\boldsymbol{e}}+\boldsymbol{i}'(\mathbf{1}-\boldsymbol{e})}\right)\leq M\right\}}\right\|$, with $f\in\mathcal{F}$, F an envelope for $\mathcal{F}$ and $M>0$.

4.7.1 Proofs of the main results

4.7.1.1 Lemma 4.3

For any $j\in\{1,...,|\mathcal{E}_r|\}$, let $L_{r,j}=\left(3|\mathcal{E}_r|^{|\mathcal{E}_r|}\right)^{|\mathcal{E}_r|-j}$. We will prove by reverse induction on j that for every function b from $\mathcal{E}_r$ to $\{1,...,|\mathcal{E}_r|\}$ with $|\mathcal{R}(b)|=j$,

$$\begin{aligned}&\mathbb{E}\Phi\left(\sup_{h\in\mathcal{H}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} h\left(\left(W^{(b(\boldsymbol{e}))}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r},\boldsymbol{i}\right)\right|\right)\\ \leq&\mathbb{E}\Phi\left(L_{r,j}\sup_{h\in\mathcal{H}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} h\left(\left(W^{(\ell(\boldsymbol{e}))}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r},\boldsymbol{i}\right)\right|\right).\end{aligned}\tag{4.16}$$

The result follows by considering $j = 1$. (4.16) is in fact an equality when $j = |\mathcal{E}_r|$, so the result holds for the base case. Next, when b is not a bijection, both sides of (4.16) are left unchanged when b is replaced by $\sigma \circ b$ for σ a permutation of $\{1, ..., |\mathcal{E}_r|\}$. As a consequence, we can assume without loss of generality that $|b^{-1}(1)| \geq 2$ and $b^{-1}(2) = ... = b^{-1}(|b^{-1}(1)|) = \emptyset$ in the induction step. This induction step is divided into two parts. In the first part, we build an array of random variables $(\widetilde{W}_A^{(e)})_{e \in \mathcal{E}_r, A \in \mathcal{A}_r}$. This array is such that

$$\begin{aligned} &\left(\left(\widetilde{W}^{(\boldsymbol{e})}_{\{\boldsymbol{i} \odot \boldsymbol{e}\}^+} \right)_{\boldsymbol{e} \in b^{-1}(1)}, \left(\widetilde{W}^{(\boldsymbol{e})}_{\{\boldsymbol{i} \odot \boldsymbol{e}\}^+} \right)_{\boldsymbol{e} \notin b^{-1}(1)} \right)_{\boldsymbol{i} \in \mathbb{I}_{n,k}} \\ \overset{d}{=} &\left(\left(W^{(\ell'(\boldsymbol{e}))}_{\{\boldsymbol{i} \odot \boldsymbol{e}\}^+} \right)_{\boldsymbol{e} \in b^{-1}(1)}, \left(W^{(b(\boldsymbol{e}))}_{\{\boldsymbol{i} \odot \boldsymbol{e}\}^+} \right)_{\boldsymbol{e} \notin b^{-1}(1)} \right)_{\boldsymbol{i} \in \mathbb{I}_{n,k}}, \end{aligned} \tag{4.17}$$

with ℓ' a bijection from $b^{-1}(1)$ to $\{1, ..., |b^{-1}(1)|\}$. Moreover, it satisfies, for all $\boldsymbol{i} \in \mathbb{I}_{n,k}$,

$$\mathbb{E}\left(h\left(\left(\widetilde{W}^{(\boldsymbol{e})}_{\{\boldsymbol{i} \odot \boldsymbol{e}\}^+} \right)_{\boldsymbol{e} \in \mathcal{E}_r}, \boldsymbol{i} \right) \middle| \mathcal{W} \right) = \frac{1}{|B(b)|} \sum_{b' \in B(b)} h\left(\left(W^{(b'(\boldsymbol{e}))}_{\{\boldsymbol{i} \odot \boldsymbol{e}\}^+} \right)_{\boldsymbol{e} \in \mathcal{E}_r}, \boldsymbol{i} \right), \tag{4.18}$$

where $\mathcal{W}$ denotes the σ-algebra generated by the $\left(W_A^{(j)}\right)_{A \in \mathcal{A}_r, j = 1, ..., |\mathcal{E}_r|}$ and

$$B(b) = \{b' : b'(\boldsymbol{e}) = b(\boldsymbol{e}) \text{ if } \boldsymbol{e} \notin b^{-1}(1),\ b'(\boldsymbol{e}) \in \{1, ..., |b^{-1}(1)|\} \text{ if } \boldsymbol{e} \in b^{-1}(1)\}.$$

In the second part of the induction step, we combine (4.17) and (4.18) with the Jensen, convexity and triangle inequalities to get upper bounds on the left-hand side of (4.16).

First part: construction of the $\widetilde{W}_A^{(\boldsymbol{e})}$.

Let ℓ' be a bijection from $b^{-1}(1)$ to $\{1, ..., |b^{-1}(1)|\}$ and let $\left(r_A^b\right)_{A \in \mathcal{A}_r}$ be some independent uniform random variables on $\{1, ..., |b^{-1}(1)|\}$. For $(j, l) \in \mathbb{N} \times \mathbb{N}^+$, rem$(j, l)$ denotes the remainder of the division of j by l. For any $(\boldsymbol{e}, A) \in \mathcal{E}_r \times \mathcal{A}_r$, let $\widetilde{W}_A^{(\boldsymbol{e})} = W_A^{\left(1 + \text{rem}(\ell'(\boldsymbol{e}) + r_A^b, |b^{-1}(1)|)\right)}$ if $\boldsymbol{e} \in b^{-1}(1)$ and $\widetilde{W}_A^{(\boldsymbol{e})} = W_A^{(b(\boldsymbol{e}))}$ otherwise. Similarly, let $\widehat{W}_A^{(\boldsymbol{e})} = W_A^{(\ell'(\boldsymbol{e}))}$ if $\boldsymbol{e} \in b^{-1}(1)$ and $\widehat{W}_A^{(\boldsymbol{e})} = W_A^{(b(\boldsymbol{e}))}$ otherwise.

Conditional on r_A^b, the function $\boldsymbol{e} \mapsto 1 + \text{rem}\left(\ell'(\boldsymbol{e}) + r_A^b, |b^{-1}(1)|\right)$ is a bijection from $b^{-1}(1)$ to $\{1, ..., |b^{-1}(1)|\}$. It follows that conditional on r_A^b, we have

$$\left(\widetilde{W}_A^{(\boldsymbol{e})} \right)_{\boldsymbol{e} \in \mathcal{E}_r} \overset{d}{=} \left(\widehat{W}_A^{(\boldsymbol{e})} \right)_{\boldsymbol{e} \in \mathcal{E}_r}.$$

Because the right-hand side does not depend on r_A^b, the previous equality also holds unconditionally. Independence of the $W_A^{(j)}$s across A ensures

$$\left(\widetilde{W}_A^{(\boldsymbol{e})} \right)_{\boldsymbol{e} \in \mathcal{E}_r, A \in \mathcal{A}_r} \overset{d}{=} \left(\widehat{W}_A^{(\boldsymbol{e})} \right)_{\boldsymbol{e} \in \mathcal{E}_r, A \in \mathcal{A}_r},$$

or equivalently

$$\left(\widetilde{W}^{(\boldsymbol{e})}_{\{\boldsymbol{i} \odot \boldsymbol{e}'\}^+} \right)_{\boldsymbol{e} \in \mathcal{E}_r, \boldsymbol{i} \in \mathbb{I}_{n,k}, \boldsymbol{e}' \in \mathcal{E}_r} \overset{d}{=} \left(\widehat{W}^{(\boldsymbol{e})}_{\{\boldsymbol{i} \odot \boldsymbol{e}'\}^+} \right)_{\boldsymbol{e} \in \mathcal{E}_r, \boldsymbol{i} \in \mathbb{I}_{n,k}, \boldsymbol{e}' \in \mathcal{E}_r}.$$

Considering elements such that $\boldsymbol{e}' = \boldsymbol{e}$ in the previous equality yields (4.17).

Next, if $(A_{\boldsymbol{e}})_{\boldsymbol{e} \in \mathcal{E}_r}$ is a family of distinct elements of $\mathcal{A}_r$, then uniform distribution and independence of the $r^b_{A_{\boldsymbol{e}}}$s induces that for every $\boldsymbol{i} \in \mathbb{I}_{n,k}$

$$\mathbb{E}\left(h\left(\left(\widetilde{W}^{(\boldsymbol{e})}_{A_{\boldsymbol{e}}} \right)_{\boldsymbol{e} \in \mathcal{E}_r}, \boldsymbol{i} \right) \middle| \mathcal{W} \right) = \frac{1}{|B(b)|} \sum_{b' \in B(b)} h\left(\left(W^{(b'(\boldsymbol{e}))}_{A_{\boldsymbol{e}}} \right)_{\boldsymbol{e} \in \mathcal{E}_r}, \boldsymbol{i} \right).$$

For every $\boldsymbol{i} \in \mathbb{I}_{n,k}$, $(\{\boldsymbol{i} \odot \boldsymbol{e}\}^+)_{\boldsymbol{e} \in \mathcal{E}_r}$ is a family of distinct subsets of $\{1, ..., n\}$ of cardinal r, so (4.18) follows.

Second part: upper bound on the LHS of (4.16)

As $\{2,...,|b^{-1}(1)|\} \cap \mathcal{R}(b) = \emptyset$, $B(b)\backslash\{b\}$ can be partitioned into two subsets $B_1(b)$ and $B_2(b)$, with

$$B_1(b) = \{b' \in B(b) : |\mathcal{R}(b')| > j = |\mathcal{R}(b)|\},$$
$$B_2(b) = \{b' \in B(b) : b'(\boldsymbol{e}) = m \in \{2,...,|b^{-1}(1)|\}\, \forall \boldsymbol{e} \in b^{-1}(1)\}.$$

Moreover, $|B_2(b)| = |b^{-1}(1)|-1$. Let $\mathcal{W}_1$ and $\mathcal{W}_1'$ be the σ-algebra generated by $\left\{W_A^{(j)}, A \in \mathcal{A}_r, j \in \mathcal{R}(b)\right\}$ and $\left\{W_A^{(j)}, A \in \mathcal{A}_r, j \in \mathcal{R}(b)\backslash\{1\}\right\}$, respectively. The $W_A^{(j)}s$ are i.i.d. across j. Consequently, for every $b' \in B_2(b)$,

$$\begin{aligned}\mathbb{E}\left(h\left(\left(W_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}^{(b'(\boldsymbol{e}))}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right)\middle|\mathcal{W}_1\right) =& \mathbb{E}\left(h\left(\left(W_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}^{(b'(\boldsymbol{e}))}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right)\middle|\mathcal{W}_1'\right)\\ =& \mathbb{E}\left(h\left(\left(W_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}^{(b(\boldsymbol{e}))}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right)\middle|\mathcal{W}_1'\right).\end{aligned}$$

As a result, using the partition $B(b) = \{b\} \cup B_1(b) \cup B_2(b)$, we obtain

$$\begin{aligned}h\left(\left(W_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}^{(b(\boldsymbol{e}))}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right) =& \mathbb{E}\left[\sum_{b'\in B(b)} h\left(\left(W_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}^{(b'(\boldsymbol{e}))}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right)\middle|\mathcal{W}_1\right]\\ &-\mathbb{E}\left[\sum_{b'\in B_1(b)} h\left(\left(W_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}^{(b'(\boldsymbol{e}))}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right)\middle|\mathcal{W}_1\right]\\ &-\left(|b^{-1}(1)|-1\right)\mathbb{E}\left[h\left(\left(W_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}^{(b(\boldsymbol{e}))}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right)\middle|\mathcal{W}_1'\right]. \qquad (4.19)\end{aligned}$$

Then, by Lemma S1.

$$\begin{aligned}&3\mathbb{E}\Phi\left(\sup_{h\in\mathcal{H}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} h\left(\left(W_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}^{(b(\boldsymbol{e}))}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right)\right|\right)\\ \leq\ & \mathbb{E}\left[\Phi\left(3\sup_{h\in\mathcal{H}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\mathbb{E}\left[\sum_{b'\in B(b)} h\left(\left(W_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}^{(b'(\boldsymbol{e}))}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right)\middle|\mathcal{W}_1\right]\right|\right)\right]\\ &+\mathbb{E}\left[\Phi\left(3\sup_{h\in\mathcal{H}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\mathbb{E}\left[\sum_{b'\in B_1(b)} h\left(\left(W_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}^{(b'(\boldsymbol{e}))}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right)\middle|\mathcal{W}_1\right]\right|\right)\right]\\ &+\mathbb{E}\left[\Phi\left(3\left(|b^{-1}(1)|-1\right)\sup_{h\in\mathcal{H}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\mathbb{E}\left[h\left(\left(W_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}^{(b(\boldsymbol{e}))}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right)\middle|\mathcal{W}_1'\right]\right|\right)\right]. \qquad (4.20)\end{aligned}$$

Denote by T_1, T_2 and T_3 the three terms on the RHS and let $\widetilde{b}(\boldsymbol{e}) = \ell'(\boldsymbol{e})$ if $\boldsymbol{e} \in b^{-1}(1)$ and $\widetilde{b}(\boldsymbol{e}) = b(\boldsymbol{e})$ otherwise. Then

$$\begin{aligned}T_1 \leq& \mathbb{E}\left[\Phi\left(3\sup_{h\in\mathcal{H}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\sum_{b'\in B(b)} h\left(\left(W_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}^{(b'(\boldsymbol{e}))}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right)\right|\right)\right]\\ =& \mathbb{E}\left[\Phi\left(3|B(b)|\sup_{h\in\mathcal{H}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\mathbb{E}\left(h\left(\left(\widetilde{W}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}^{(\boldsymbol{e})}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right)\middle|\mathcal{W}\right)\right|\right)\right]\\ \leq& \mathbb{E}\left[\Phi\left(3|B(b)|\sup_{h\in\mathcal{H}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} h\left(\left(W_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}^{(\widetilde{b}(\boldsymbol{e}))}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right)\right|\right)\right]\\ \leq& \mathbb{E}\left[\Phi\left(3|B(b)|L_{r,j+1}\sup_{h\in\mathcal{H}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} h\left(\left(W_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}^{(\ell(\boldsymbol{e}))}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right)\right|\right)\right]. \qquad (4.21)\end{aligned}$$

The first inequality follows by Jensen's inequality. The first equality is due to (4.18). The second inequality uses Jensen's inequality and (4.17). Finally, (4.21) relies on the induction hypothesis and $|\mathcal{R}(\widetilde{b})| > j$. Similarly,

$$\begin{aligned} T_2 \leq & \frac{1}{|B_1(b)|} \sum_{b' \in B_1(b)} \mathbb{E}\left[\Phi\left(3|B_1(b)| \sup_{h\in\mathcal{H}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} h\left(\left(W^{(b'(\boldsymbol{e}))}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right)\right|\right)\right] \\ \leq & \mathbb{E}\left[\Phi\left(3|B_1(b)|L_{r,j+1} \sup_{h\in\mathcal{H}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} h\left(\left(W^{(\ell(\boldsymbol{e}))}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right)\right|\right)\right], \end{aligned} \tag{4.22}$$

where the first inequality follows by Jensen's inequality and the second by the induction hypothesis, since $|\mathcal{R}(b')| > j$ for all $b' \in B_1(b)$. Finally, note that for each $\boldsymbol{i}$, all the $\{\boldsymbol{i}\odot\boldsymbol{e}\}^+$s are disjoint so, conditional on $\mathcal{W}_1'$,

$$\left(W^{(b(\boldsymbol{e}))}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r} \overset{d}{=} \left(W^{(\widetilde{b}(\boldsymbol{e}))}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r}.$$

As a result,

$$\begin{aligned} T_3 = & \mathbb{E}\left[\Phi\left(3\left(|b^{-1}(1)|-1\right) \sup_{h\in\mathcal{H}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} \mathbb{E}\left[h\left(\left(W^{(\widetilde{b}(\boldsymbol{e}))}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right)\middle|\mathcal{W}_1'\right]\right|\right)\right] \\ \leq & \mathbb{E}\left[\Phi\left(3\left(|b^{-1}(1)|-1\right) L_{r,j+1} \sup_{h\in\mathcal{H}}\left|\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} h\left(\left(W^{(\ell(\boldsymbol{e}))}_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\right)_{\boldsymbol{e}\in\mathcal{E}_r}, \boldsymbol{i}\right)\right|\right)\right], \end{aligned} \tag{4.23}$$

where the inequality follows by Jensen's inequality and the induction hypothesis again. We finally get (4.16) by combining (4.20)-(4.23) with monotonicity of Φ, the expression of $L_{r,j+1}$ and

$$\max\left(|B(b)|, |B_1(b)|, |b^{-1}(1)-1|\right) \leq |\mathcal{E}_r|^{|\mathcal{E}_r|}.$$

This concludes the induction step, and thus the proof of the lemma.

4.7.1.2 Theorem 4.1

4.7.1.2.1 Uniform law of large numbers

Convergence in L^1. Let M be some arbitrary positive constant. The symmetrization Lemma 4.2 applied to the class $\mathcal{G} = \left\{f\mathbb{1}_{\{F\leq M\}}, f\in\mathcal{F}\right\}$ and $\Phi = \mathrm{Id}$ ensures that

$$\begin{aligned} \mathbb{E}\left[\sup_{\mathcal{F}}|\mathbb{P}_n f - Pf|\right] \leq & 2\mathbb{E}\left[F\left(Y_{\mathbf{1}}\right)\mathbb{1}_{\{F(Y_{\mathbf{1}})>M\}}\right] \\ & + \sum_{r=1}^{k}\sum_{\boldsymbol{e}\in\mathcal{E}_r} K_{r,k}\mathbb{E}\left[\sup_{\mathcal{F}}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} \varepsilon_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+} f\left(Y^r_{\boldsymbol{i}}\right)\mathbb{1}_{\{F(Y^r_{\boldsymbol{i}})\leq M\}}\right|\right], \end{aligned}$$

with $K_{r,k}$ some non negative number depending on r and k only.

For every $(a_{ij})_{i=1\ldots n, j=1\ldots m} \in \mathbb{R}^{nm}$ and independent Rademacher random variables $(\varepsilon_i)_{i=1\ldots n}$, we have [see for instance Lemma 2.3.4 in 79]

$$\mathbb{E}\left[\max_{j\in\{1,\ldots,m\}}\left|\sum_{i=1}^{n}\varepsilon_i a_{ij}\right|\right] \leq \left[2\log(2m)\max_{j\in\{1,\ldots,m\}}\sum_{i=1}^{n}a_{ij}^2\right]^{1/2}. \tag{4.24}$$

Next, reasoning conditionally on the data, we can consider for every $\eta_1 > 0$ a minimal η_1-covering of $\mathcal{F}$ for

the seminorm $||\cdot||_{\boldsymbol{e},M,1}$ with closed balls centered in $\mathcal{F}$. This implies

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathcal{F}}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}f\left(Y_{\boldsymbol{i}}^r\right)\mathbb{1}_{\{F(Y_{\boldsymbol{i}}^r)\leq M\}}\right|\middle|(Y_{\boldsymbol{i}}^r)_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\right]\\
=&\mathbb{E}\left[\sup_{\mathcal{F}}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\overrightarrow{\mathbb{I}_{n,r}}}\varepsilon_{\{\boldsymbol{i}\}}\sum_{\pi\in\mathfrak{S}_r}\sum_{\boldsymbol{i}'\in\overline{(\{1,\dots,n\}\setminus\{\boldsymbol{i}\})^{k-r}}}f\left(Y^r_{(\boldsymbol{i}_\pi)^{\boldsymbol{e}}+\boldsymbol{i}'^{(\mathbf{1}-\boldsymbol{e})}}\right)\mathbb{1}_{\left\{F\left(Y^r_{(\boldsymbol{i}_\pi)^{\boldsymbol{e}}+\boldsymbol{i}'^{(\mathbf{1}-\boldsymbol{e})}}\right)\leq M\right\}}\right|\middle|(Y_{\boldsymbol{i}}^r)_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\right]\\
\leq&M\left(\frac{2\log 2N\left(\eta_1,\mathcal{F},||.||_{\boldsymbol{e},M,1}\right)(n-r)!r!}{n!}\right)^{1/2}+\eta_1. \qquad (4.25)
\end{aligned}$$

To obtain the inequality, we apply (4.24) with $m=N\left(\eta_1,\mathcal{F},||.||_{\boldsymbol{e},M,1}\right)$ and

$$a_{\boldsymbol{i}j}=\frac{(n-k)!}{n!}\sum_{\pi\in\mathfrak{S}_r}\sum_{\boldsymbol{i}'\in\overline{(\{1,\dots,n\}\setminus\{\boldsymbol{i}\})^{k-r}}}f_j\left(Y^r_{(\boldsymbol{i}_\pi)^{\boldsymbol{e}}+\boldsymbol{i}'^{(\mathbf{1}-\boldsymbol{e})}}\right)\mathbb{1}_{\{F(Y^r_{(\boldsymbol{i}_\pi)^{\boldsymbol{e}}+\boldsymbol{i}'^{(\mathbf{1}-\boldsymbol{e})}})\leq M\}},$$

where f_j is one of the $N\left(\eta_1,\mathcal{F},||.||_{\boldsymbol{e},M,1}\right)$ centers of balls needed to cover $\mathcal{F}$. Inequality then (4.25) follows by remarking that

$$\left(\sum_{i=1}^n a_{\boldsymbol{i}j}^2\right)^{1/2}\leq M\binom{n}{r}^{1/2}\frac{(n-k)!}{n!}r!\frac{(n-r)!}{(n-r-(k-r))!}=M\left(\frac{(n-r)!r!}{n!}\right)^{1/2}.$$

Observe that $||g||_{\boldsymbol{e},M,1}\leq||g||_{\mathbb{Q}_n^r,1}$. Thus, considering $\eta_1=\eta||F||_{\mathbb{Q}_n^r,1}$ and using Point 2 of Lemma S4.11, we have, for every $\eta>0$,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathcal{F}}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}f\left(Y_{\boldsymbol{i}}^r\right)\mathbb{1}_{\{F(Y_{\boldsymbol{i}}^r)\leq M\}}\right|\middle|(Y_{\boldsymbol{i}}^r)_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\right]\\
\leq&M\left(\frac{2\log 2\sup_Q N\left(\eta||F||_{Q,1},\mathcal{F},||.||_{Q,1}\right)(n-r)!r!}{n!}\right)^{1/2}+\eta||F||_{\mathbb{Q}_n^r,1}.
\end{aligned}$$

For any r and any $\boldsymbol{i}\in\mathbb{I}_k$, we have $\mathbb{E}(F(Y_{\boldsymbol{i}}^r))=\mathbb{E}(F(Y_{\mathbf{1}}))$, and next $\mathbb{E}\left(||F||_{\mathbb{Q}_n^r,1}\right)=\mathbb{E}(F(Y_{\mathbf{1}}))$. Integration with respect to the distribution of $(Y_{\boldsymbol{i}}^r)_{\boldsymbol{i}\in\mathbb{I}_{n,k}}$ ensures

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathcal{F}}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}f\left(Y_{\boldsymbol{i}}^r\right)\mathbb{1}_{\{F(Y_{\boldsymbol{i}}^r)\leq M\}}\right|\right]\\
\leq&M\left(\frac{2\log 2\sup_Q N\left(\eta||F||_{Q,1},\mathcal{F},||.||_{Q,1}\right)(n-r)!r!}{n!}\right)^{1/2}+\eta\mathbb{E}\left(F(Y_{\mathbf{1}})\right).
\end{aligned}$$

It follows that there exists a constant K_k' such that

$$\begin{aligned}
\mathbb{E}\left[\sup_{\mathcal{F}}|\mathbb{P}_nf-Pf|\right]\leq K_k'\Bigg(&\mathbb{E}\left[F\left(Y_{\mathbf{1}}\right)\mathbb{1}_{\{F(Y_{\mathbf{1}})>M\}}\right]\\
&+M\left(\frac{2\log 2\sup_Q N\left(\eta||F||_{Q,1},\mathcal{F},||.||_{Q,1}\right)}{n}\right)^{1/2}+\eta\mathbb{E}(F(Y_{\mathbf{1}}))\Bigg).
\end{aligned}$$

Picking M and η such that $\mathbb{E}\left[F\left(Y_{\mathbf{1}}\right)\mathbb{1}_{\{F(Y_{\mathbf{1}})>M\}}\right]+\eta\mathbb{E}(F(Y_{\mathbf{1}}))$ is small and letting n tend to infinity, we conclude that $\mathbb{E}\left[\sup_{\mathcal{F}}|\mathbb{P}_nf-Pf|\right]=o(1)$.

Almost-sure convergence. Let Σ_n the σ-algebra generated by $\mathcal{H}_n$, the set of functions g from $\mathcal{Y}^{\mathbb{I}_k}$ to $\mathbb{R}$ that are invariant by the action of any permutation π on $\mathbb{N}^+$ such that $\pi(j)=j$ for $j\geq n$:

$$g\left((Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)=g\left((Y_{(\pi(i_1),\dots,\pi(i_k))})_{\boldsymbol{i}\in\mathbb{I}_k}\right).$$

Let $h\left((Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\right)=\sup_{\mathcal{F}}|\mathbb{P}_n f-Pf|$ and for $l=1,...,n+1$, let $\mathbb{P}_{n+1}^{\backslash\{l\}}f=\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n+1,k}}f\left(Y_{\boldsymbol{i}}\right)\mathbb{1}_{\{l\notin\{\boldsymbol{i}\}\}}$. Let π denote the transposition on $\mathbb{N}^+$ exchanging $n+1$ and l. Exchangeability and the definition of $\mathcal{H}_n$ ensure that

$$\begin{aligned}&\left((Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\overline{(\{1,...,n+1\}\backslash\{l\})^k}},\left(g\left((Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)\right)_{g\in\mathcal{H}_{n+1}}\right)\\ \overset{d}{=}&\left((Y_{\pi(\boldsymbol{i})})_{\boldsymbol{i}\in\overline{(\{1,...,n+1\}\backslash\{l\})^k}},\left(g\left(\left(Y_{\pi(\boldsymbol{i})}\right)_{\boldsymbol{i}\in\mathbb{I}_k}\right)\right)_{g\in\mathcal{H}_{n+1}}\right)\\ \overset{a.s.}{=}&\left((Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_{n,k}},\left(g\left((Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)\right)_{g\in\mathcal{H}_{n+1}}\right).\end{aligned}$$

For every $l<n+1$, the above implies that conditional on Σ_{n+1},

$$(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\overline{(\{1,...,n+1\}\backslash\{l\})^k}}\overset{d}{=}(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_{n,k}}.$$

As a result,

$$\begin{aligned}\mathbb{E}\left(\sup_{\mathcal{F}}\left|\mathbb{P}_{n+1}^{\backslash\{l\}}f-Pf\right|\Big|\Sigma_{n+1}\right)&=\mathbb{E}\left(h((Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\overline{(\{1,...,n+1\}\backslash\{l\})^k}})\Big|\Sigma_{n+1}\right)\\&=\mathbb{E}\left(h((Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_{n,k}})\Big|\Sigma_{n+1}\right)\\&=\mathbb{E}\left(\sup_{\mathcal{F}}|\mathbb{P}_n f-Pf|\Big|\Sigma_{n+1}\right).\end{aligned}$$

Because $\sum_{l=1}^{n+1}\mathbb{P}_{n+1}^{\backslash\{l\}}f=\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n+1,k}}\sum_{l=1}^{n+1}f(Y_{\boldsymbol{i}})\mathbb{1}_{\{l\notin\{\boldsymbol{i}\}\}}=\frac{(n+1-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n+1,k}}f(Y_{\boldsymbol{i}})$, we have

$$\frac{1}{n+1}\sum_{l=1}^{n+1}\mathbb{P}_{n+1}^{\backslash\{l\}}f=\mathbb{P}_{n+1}f.$$

The triangle inequality ensures

$$\sup_{\mathcal{F}}|\mathbb{P}_{n+1}f-Pf|\leq\frac{1}{n+1}\sum_{l=1}^{n+1}\sup_{\mathcal{F}}\left|\mathbb{P}_{n+1}^{\backslash\{l\}}f-Pf\right|.$$

Combining the last inequality with $\mathbb{E}\left(\sup_{\mathcal{F}}|\mathbb{P}_{n+1}f-Pf|\Big|\Sigma_{n+1}\right)=\sup_{\mathcal{F}}|\mathbb{P}_{n+1}f-Pf|$, we finally obtain

$$\begin{aligned}\sup_{\mathcal{F}}|\mathbb{P}_{n+1}f-Pf|&\leq\frac{1}{n+1}\sum_{l=1}^{n+1}\mathbb{E}\left(\sup_{\mathcal{F}}\left|\mathbb{P}_{n+1}^{\backslash\{l\}}f-Pf\right|\Big|\Sigma_{n+1}\right)\\&=\mathbb{E}\left(\sup_{\mathcal{F}}|\mathbb{P}_n f-Pf|\Big|\Sigma_{n+1}\right).\end{aligned}$$

This means that $\sup_{\mathcal{F}}|\mathbb{P}_n f-Pf|$ is a reverse submartingale with respect to the decreasing filtration Σ_n. Hence, by the convergence theorem for backwards submartingale [see, e.g., Theorem 22 of Chapter 24 in 73] and its convergence to 0 in L^1, $\sup_{\mathcal{F}}|\mathbb{P}_n f-Pf|$ converges almost surely to 0.

4.7.1.2.2 Uniform central limit theorem

To prove this result, we follow a usual strategy which consists in showing the pointwise convergence, asymptotic equicontinuity and total boundedness of $\mathcal{F}$ [see for instance, 137].

First step: pointwise convergence

Let $(f_1,...,f_m)\in\mathcal{F}\times...\times\mathcal{F}$. The Cramer-Wold device ensures the joint asymptotic normality of $(f_1,...,f_m)$ if the asymptotic normality holds for $f=\sum_{i=1}^m\lambda_i f_i$ for every $(\lambda_1,...,\lambda_m)\in\mathbb{R}^m$ such that $\sum_{i=1}^m|\lambda_i|<+\infty$.

For $f \in L^2(P)$, $\widehat{\theta} = \frac{(n-k)!}{n!} \sum_{\boldsymbol{i} \in \mathbb{I}_{n,k}} f(Y_{\boldsymbol{i}})$ denotes the estimator of $\theta_0 = \mathbb{E}(f(Y_{\mathbf{1}}))$. Theorem A in [128] ensures that

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow[n \to +\infty]{d} \mathcal{N}(0, K(f,f)).$$

Second step: asymptotic equicontinuity

We have to show that, for every $\epsilon > 0$

$$\lim_{\delta \to 0} \limsup_{n \to +\infty} \mathbb{P}\left(\sup_{f \in \mathcal{F}_\delta} |\mathbb{G}_n f| > \epsilon\right) = 0. \tag{4.26}$$

By Markov's inequality, it is sufficient to show

$$\lim_{\delta \to 0} \limsup_{n \to +\infty} \mathbb{E}\left[\sup_{\mathcal{F}_\delta} |\mathbb{G}_n f|\right] = 0. \tag{4.27}$$

A weighted Rademacher empirical process is sub-Gaussian with respect to the Euclidean norm of the vector of weights. As a result, conditionally on the original data, we can apply Theorem 2.3.6 in [79]. This observation implies that for every $r = 1, ..., k$ and $\boldsymbol{e} \in \mathcal{E}_r$,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathcal{F}_\delta} \left| \frac{(n-k)!}{n!} \sum_{\boldsymbol{i} \in \mathbb{I}_{n,k}} \varepsilon_{\{\boldsymbol{i} \odot \boldsymbol{e}\}^+} f\left(Y^r_{\boldsymbol{i}}\right) \right| \middle| (Y^r_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{I}_{n,k}} \right] \\
=&\mathbb{E}\left[\sup_{\mathcal{F}_\delta} \left| \frac{(n-k)!}{n!} \sum_{\boldsymbol{i} \in \overrightarrow{\mathbb{I}_{n,r}}} \varepsilon_{\{\boldsymbol{i}\}} \sum_{\pi \in \mathfrak{S}_r} \sum_{\boldsymbol{i}' \in \overrightarrow{(\{1,...,n\} \setminus \{\boldsymbol{i}\})^{k-r}}} f\left(Y^r_{(\boldsymbol{i}_\pi)^{\boldsymbol{e}} + \boldsymbol{i}'^{(\mathbf{1}-\boldsymbol{e})}}\right) \right| \middle| (Y^r_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{I}_{n,k}} \right] \\
\leq& \frac{4\sqrt{2(n-r)!r!}}{\sqrt{n!}} \int_0^{\sigma_{\boldsymbol{e}}} \sqrt{\log 2N\left(\varepsilon, \mathcal{F}_\delta, ||\cdot||_{\boldsymbol{e},2}\right)} d\varepsilon,
\end{aligned}$$

with

$$||f||^2_{\boldsymbol{e},2} = \frac{(n-r)!r!}{n!} \sum_{\boldsymbol{i} \in \overrightarrow{\mathbb{I}_{n,r}}} \left(\frac{(n-k)!}{(n-r)!r!} \sum_{\pi \in \mathfrak{S}_r} \sum_{\boldsymbol{i}' \in \overrightarrow{(\{1,...,n\} \setminus \{\boldsymbol{i}\})^{k-r}}} f\left(Y^r_{(\boldsymbol{i}_\pi)^{\boldsymbol{e}} + \boldsymbol{i}'^{(\mathbf{1}-\boldsymbol{e})}}\right) \right)^2$$

and $\sigma^2_{\boldsymbol{e}} = \sup_{\mathcal{F}_\delta} ||f||^2_{\boldsymbol{e},2}$. A convexity argument ensures $||f||^2_{\boldsymbol{e},2} \leq ||f||^2_{\mathbb{P}^r_n,2}$. As a result, $N\left(\varepsilon, \mathcal{F}_\delta, ||\cdot||_{\boldsymbol{e},2}\right) \leq N\left(\varepsilon, \mathcal{F}_\delta, ||\cdot||_{\mathbb{Q}^r_n,2}\right)$ and $\sigma^2_{\boldsymbol{e}} \leq \sigma^2_r$, with $\sigma^2_r = \sup_{\mathcal{F}_\delta} ||f||^2_{\mathbb{Q}^r_n,2}$. Next for every $r = 1, ..., k$ and $\boldsymbol{e} \in \mathcal{E}_r$:

$$\begin{aligned}
&\sqrt{n}\mathbb{E}\left[\sup_{\mathcal{F}_\delta} \left| \frac{(n-k)!}{n!} \sum_{\boldsymbol{i} \in \mathbb{I}_{n,k}} \varepsilon_{\{\boldsymbol{i} \odot \boldsymbol{e}\}^+} f\left(Y^r_{\boldsymbol{i}}\right) \right| \middle| (Y^r_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{I}_{n,k}} \right] \\
\leq& 4\sqrt{2k!} \int_0^{\sigma_r} \sqrt{\log 2N\left(\varepsilon, \mathcal{F}_\delta, ||\cdot||_{\mathbb{Q}^r_n,2}\right)} d\varepsilon.
\end{aligned}$$

Lemma 4.2 applied to the class $\mathcal{F}_\delta$ then implies

$$\mathbb{E}\left[\sup_{\mathcal{F}_\delta} |\mathbb{G}_n f|\right] = \sum_{r=1}^k O\left(\mathbb{E}\left(\int_0^{\sigma_r} \sqrt{\log 2N\left(\varepsilon, \mathcal{F}_\delta, ||\cdot||_{\mathbb{Q}^r_n,2}\right)} d\varepsilon\right)\right).$$

Since $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $\mathbb{Q}^r_n$ is a (random) probability measure on $\mathcal{Y}$ with finite support for any $r = 1, ..., k$, we obtain

$$\begin{aligned}
&\int_0^{\sigma_r} \sqrt{\log 2N\left(\varepsilon, \mathcal{F}_\delta, ||\cdot||_{\mathbb{Q}^r_n,2}\right)} d\varepsilon \\
\leq& \sqrt{\log 2}\sigma_r + ||F||_{\mathbb{Q}^r_n,2} \int_0^{\sigma_r / ||F||_{\mathbb{Q}^r_n,2}} \sup_Q \sqrt{\log N(\eta ||F||_{Q,2}, \mathcal{F}_\delta, ||.||_{Q,2})} d\eta.
\end{aligned}$$

Let $J_{\mathcal{F}_\delta}(u) = \int_0^u \sup_Q \sqrt{\log N(\eta||F||_{Q,2}, \mathcal{F}_\delta, ||.||_{Q,2})} d\eta$. The functions $x \mapsto \sqrt{x}$ and $(x,y) \mapsto \sqrt{y} J_{\mathcal{F}_\delta}(\sqrt{x}/\sqrt{y})$ are both concave (the latter in view of Point 2 of Lemma S4.10) and $\mathbb{E}\left(||F||^2_{\mathbb{P}^r_n,2}\right) = \mathbb{E}\left(||F^2||_{\mathbb{P}^r_n,1}\right) = \mathbb{E}(F^2(Y_{\mathbf{1}}))$. Then, by Jensen's inequality,

$$\mathbb{E}\left[\sup_{\mathcal{F}_\delta} |\mathbb{G}_n f|\right] = \sum_{r=1}^{k} O\left(\mathbb{E}(\sigma_r^2)^{1/2} + \mathbb{E}(F^2(Y_{\mathbf{1}}))^{1/2} J_{\mathcal{F}_\delta}\left(\frac{\mathbb{E}(\sigma_r^2)^{1/2}}{\mathbb{E}(F^2(Y_{\mathbf{1}}))^{1/2}}\right)\right).$$

Thanks to Points 3 and 4 of Lemmas S4.11, we further get

$$\mathbb{E}\left[\sup_{\mathcal{F}_\delta} |\mathbb{G}_n f|\right] = \sum_{r=1}^{k} O\left(\mathbb{E}(\sigma_r^2)^{1/2} + \mathbb{E}(F^2(Y_{\mathbf{1}}))^{1/2} J_{\mathcal{F}}\left(\frac{\mathbb{E}(\sigma_r^2)^{1/2}}{4\mathbb{E}(F^2(Y_{\mathbf{1}}))^{1/2}}\right)\right).$$

As $\lim_{x\downarrow 0} J_{\mathcal{F}}(x) = 0$ and $J_{\mathcal{F}}$ and $x \mapsto \sqrt{x}$ are non-decreasing, it is sufficient to show that

$$\lim_{\delta\downarrow 0} \limsup_{n\to+\infty} \mathbb{E}(\sigma_r^2) = 0, \text{ for every } r = 1, ..., k \tag{4.28}$$

By the triangle inequality and the definition of $\mathcal{F}_\delta$ and $\mathcal{F}_\infty$,

$$\begin{aligned}\sigma_r^2 = \sup_{\mathcal{F}_\delta} \left|\frac{(n-k)!}{n!} \sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} f^2(Y_{\boldsymbol{i}}^r)\right| &\leq \sup_{\mathcal{F}_\delta} \left|\frac{(n-k)!}{n!} \sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} f^2(Y_{\boldsymbol{i}}^r) - Pf^2\right| + \delta^2 \\ &\leq \sup_{\mathcal{F}_\infty} \left|\frac{(n-k)!}{n!} \sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} f^2(Y_{\boldsymbol{i}}^r) - Pf^2\right| + \delta^2.\end{aligned}$$

Noting that $4F^2$ is an envelope for $\mathcal{F}_\infty^2$, Point 5 of Lemma S4.11 yields

$$\sup_Q N\left(\eta||4F^2||_{Q,1}, \mathcal{F}_\infty^2, ||.||_{Q,1}\right) < +\infty \text{ for every } \eta > 0.$$

Applying Theorem 4.1 to the class $\mathcal{F}_\infty^2$ for the array $(Y_{\boldsymbol{i}}^r)_{\boldsymbol{i}\in\mathbb{I}_k}$, we get

$$\lim_{n\to\infty} \mathbb{E}\left(\sup_{\mathcal{F}_\infty} \left|\frac{(n-k)!}{n!} \sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} f^2(Y_{\boldsymbol{i}}^r) - Pf^2\right|\right) = 0,$$

and then (4.28) holds.

Third step: total boundedness

Fix $\varepsilon > 0$. The reasoning previously used to control σ_r ensures $\lim_{n\to\infty} \mathbb{E}\left(\sup_{\mathcal{F}_\infty} \left|\mathbb{P}_n f^2 - Pf^2\right|\right) = 0$. Then we have with probability approaching one and for every $(f_1, f_2) \in \mathcal{F} \times \mathcal{F}$

$$||f_1 - f_2||^2_{P,2} \leq ||f_1 - f_2||^2_{\mathbb{P}_n,2} + \varepsilon^2.$$

As a consequence,

$$\begin{aligned}N(\varepsilon, \mathcal{F}, ||.||_{P,2}) &\leq N\left(\frac{\varepsilon}{\sqrt{2}}, \mathcal{F}, ||.||_{\mathbb{P}_n,2}\right) + o_p(1) \\ &\leq \mathbb{1}_{\{||F||_{\mathbb{P}_n,2}=0\}} + \sup_Q N\left(\frac{\varepsilon||F||_{Q,2}}{\sqrt{2}||F||_{\mathbb{P}_n,2}}, \mathcal{F}, ||.||_{Q,2}\right) \mathbb{1}_{\{||F||_{\mathbb{P}_n,2}>0\}} + o_p(1) = O_p(1),\end{aligned}$$

because $||F||_{\mathbb{P}_n,2}$ converges almost-surely to $\mathbb{E}(F^2(Y_{\mathbf{1}}))^{1/2}$ and then $N(\varepsilon, \mathcal{F}, ||.||_{P,2}) < +\infty$.

4.7.1.3 Theorem 4.2

We only have to prove the pointwise convergence and the asymptotic equicontinuity, since the total boundedness of $\mathcal{F}$ is already proved in Theorem 4.1.

For the boostrap, we sample n units independently in $\{1,...,n\}$ with replacement and equal probability. For $i=1,...,n$, i^* denotes the i-th sampled unit and for $\boldsymbol{i}\in\mathbb{I}_{n,k}$, $\boldsymbol{i}^*$ denotes $(i_1^*,...,i_k^*)$. We then have: $\mathbb{P}_n^* f=\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} f(Y_{\boldsymbol{i}^*})\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}$.

First step: pointwise convergence

Let $(f_1,...,f_m)\in\mathcal{F}\times...\times\mathcal{F}$. To prove convergence of the bootstrap for the finite subclass $(f_1,...,f_m)$, the Cramer-Wold device ensures it is sufficient to prove the asymptotic normality for $f=\sum_j\lambda_j f_j$ and every $(\lambda_1,...,\lambda_m)\in\mathbb{R}^m$.

Substep 1: asymptotic equivalence

Let $\theta=\mathbb{E}(f(Y_{\mathbf{1}}))$ the parameter of interest, $\theta^*=\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} f(Y_{\boldsymbol{i}^*})\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}$ its estimator and $\widehat{\theta}=\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} f(Y_{\boldsymbol{i}})$ the corresponding bootstrap estimator. For $\boldsymbol{i}\in\{1,...,n\}^k$, let $h(\boldsymbol{i})=\mathbb{1}_{\{\boldsymbol{i}\in\mathbb{I}_{n,k}\}}\sum_{\pi\in\mathfrak{S}_k} f(Y_{\boldsymbol{i}_\pi})$. We have $\theta^*=\frac{(n-k)!}{n!k!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} h(\boldsymbol{i}^*)$, $\widehat{\theta}=\frac{(n-k)!}{n!k!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} h(\boldsymbol{i})$ $=\frac{(n-k)!}{n!k!}\sum_{\boldsymbol{i}\in\{1,...,n\}^k} h(\boldsymbol{i})$ and $\mathbb{E}\left(\theta^*|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)=\frac{n!}{n^k(n-k)!}\widehat{\theta}$. Let

$$\theta_1^*=\frac{(n-k)!}{n!k!}\sum_{\boldsymbol{j}\in\{1,...,n\}^k} h(j_1^*,j_2,...,j_k).$$

We have $\mathbb{E}\left(\theta_1^*|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)=\widehat{\theta}$. For $(\boldsymbol{i},\boldsymbol{j})\in\mathbb{I}_{n,k}\times\{1,...,n\}^k$, observe that

$$\begin{aligned}&\mathbb{E}\left(h(\boldsymbol{i}^*)h(j_1^*,j_2,...,j_k)|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)\\ &=\begin{cases}\frac{n!k!}{n^k(n-k)!}\widehat{\theta}\times\frac{1}{n}\sum_{j=1}^n h(j,j_2,...,j_k) & \text{if } j_1\notin\{i_1,...,i_k\}\\ \frac{1}{n^k}\sum_{i_1=1}^n\left(\sum_{(i_2,...,i_k)\in\{1,...,n\}^{k-1}} h(\boldsymbol{i})\times h(i_1,j_2,...,j_k)\right) & \text{otherwise.}\end{cases}\end{aligned}$$

Consequently,

$$\begin{aligned}n\mathbb{E}\left(\theta^*\theta_1^*|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)=&n\frac{(n-k)!^2}{n!^2k!^2}(n-k)\frac{n!}{(n-k)!}\left(\frac{n!k!}{n^k(n-k)!}\widehat{\theta}\frac{1}{n}\sum_{\boldsymbol{j}\in\mathbb{I}_{n,k}} h(\boldsymbol{j})\right)\\ &+n\frac{(n-k)!^2}{n!^2k!^2}k\frac{n!}{(n-k)!}\frac{1}{n^k}\sum_{i_1=1}^n\left(\sum_{(i_2,...,i_k)\in\{1,...,n\}^{k-1}} h(\boldsymbol{i})\right)^2\\ =&\frac{n-k}{n^k}\frac{n!}{(n-k)!}\widehat{\theta}^2\\ &+\frac{k}{k!^2}\frac{n^k(n-k)!}{n!}\frac{1}{n^{2k-1}}\sum_{i_1=1}^n\left(\sum_{(i_2,...,i_k)\in\{1,...,n\}^{k-1}} h(\boldsymbol{i})\right)^2.\end{aligned}$$

Focusing on the last sum, Lemma S4.8 allows us to conclude that

$$\begin{aligned}&\sum_{i_1=1}^n\left(\sum_{(i_2,...,i_k)\in\{1,...,n\}^{k-1}} h(\boldsymbol{i})\right)^2\\ =&\sum_{\boldsymbol{j}\in\{1,...,n\}^{2k-1}} h(j_1,...,j_k)h(j_1,j_{k+1},...,j_{2k-1})\\ =&\sum_{c=0}^{k-1}\binom{k-1}{c}^2\left(n^{2k-1-c}\mathbb{E}\left[h(1,...,k)h(1,...,1+c,k+1,...,2k-c-1)\right]+o_{a.s}(n^{2k-1-c})\right).\end{aligned}$$

As $\frac{n^k(n-k)!}{n!}$ converges to 1, the quantity $\frac{n^k(n-k)!}{n!}\frac{1}{n^{2k-1}}\sum_{i_1=1}^n\left(\sum_{(i_2,...,i_k)\in\{1,...,n\}^{k-1}} h(\boldsymbol{i})\right)^2$ converges almost surely to $\mathbb{E}\left(h(\mathbf{1})h(\mathbf{1}')\right)$.

Combining the exchangeability assumption, symmetry of h and a combinatorial argument [see the

proof of Theorem 12.3 in 136], we obtain

$$\begin{aligned}
&n\mathbb{E}\left(\theta^{*2}|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)\\
=&n\frac{(n-k)!^2}{n!^2k!^2}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\sum_{\boldsymbol{j}\in\mathbb{I}_{n,k}}\mathbb{E}\left(h(\boldsymbol{i}^*)h(\boldsymbol{j}^*)\big|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)\\
=&n\frac{(n-k)!}{n!k!^2}\mathbb{E}\left(h(1^*,...,k^*)^2\big|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)\\
=&n\frac{(n-k)!^2}{n!^2}\sum_{l=0}^{k}\binom{n}{k}\binom{k}{l}\binom{n-k}{k-l}\mathbb{E}\left[h(1^*,...,k^*)h(1^*,...,l^*,(k+1)^*,...,(2k-l)^*)\big|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right].
\end{aligned}$$

When $l=0$

$$\begin{aligned}
&n\frac{(n-k)!^2}{n!^2}\binom{n}{k}\binom{k}{l}\binom{n-k}{k-l}\mathbb{E}\left[h(1^*,...,k^*)h(1^*,...,l^*,(k+1)^*,...,(2k-l)^*)\big|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right]\\
=&\frac{n(n-k)!^2n!}{n!^2k!^2(n-2k)!}\left(\frac{1}{n^k}\sum_{\boldsymbol{i}\in\{1,...,n\}^k}h(\boldsymbol{i})\right)^2=n\frac{n!}{n^{2k}(n-2k)!}\widehat{\theta}^2.
\end{aligned}$$

For every $l=1,...,k$,

$$\begin{aligned}
&\mathbb{E}\left[h(1^*,...,(k-1)^*,k^*)h(1^*,...,l^*,(k+1)^*,...,(2k-l)^*)\big|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right]\\
=&\frac{1}{n^l}\sum_{\boldsymbol{i}\in\overline{\{1,...,n\}^l}}\left(\frac{1}{n^{k-l}}\sum_{\boldsymbol{j}\in\overline{\{1,...,n\}}^{k-l}}h(i_1,...,i_l,j_1,...,j_{k-l})\right)^2\\
=&\frac{1}{n^{2k-l}}\sum_{\boldsymbol{j}\in\{1,...,n\}^{2k-l}}h(j_1,...,j_k)h(j_1,...,j_l,j_{k+1},...,j_{2k-l})\\
=&\frac{1}{n^{2k-l}}\sum_{c=0}^{k-l}\binom{k-l}{c}^2\left(n^{2k-l-c}\mathbb{E}\left[h(1,...,k)h(1,...,l+c,k+1,...,2k-c-l)\right]+o_{a.s}(n^{2k-l-c})\right)\\
=&\mathbb{E}\left[h(1,...,k)h(1,...,l,k+1,...,2k-l)\right]+o_{a.s}(1),
\end{aligned}$$

using Lemma S4.8 once more. As $n\frac{(n-k)!^2}{n!^2}\binom{n}{k}\binom{k}{l}\binom{n-k}{k-l}=O(n^{1-k+k-l})=o(1)$ for every $l\geq 2$ and $n\frac{(n-k)!^2}{n!^2}\binom{n}{k}\binom{k}{1}\binom{n-k}{k-1}=\frac{k^2}{k!^2}+o(1)$, we get

$$\begin{aligned}
&n\mathbb{E}\left(\theta^{*2}|(Y_{\boldsymbol{i}'})^k_{\boldsymbol{i}'\in\overline{\{1,...,n\}}}\right)\\
=&n\frac{n!}{n^{2k}(n-2k)!}\widehat{\theta}^2+\frac{k^2}{k!^2}\mathbb{E}\left[h(1,...,k)h(1,...,l,k+1,...,2k-1)\right]+o_{a.s}(1).
\end{aligned}$$

We also have

$$\begin{aligned}
n\mathbb{E}\left(\theta_1^{*2}\big|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)&=n\frac{(n-k)!^2}{n!^2k!^2}\sum_{\boldsymbol{i}\in\{1,...,n\}^k}\sum_{\boldsymbol{j}\in\{1,...,n\}^k}\mathbb{E}\left(h(i_1^*,i_2,...,i_k)h(j_1^*,j_2,...,j_k)\big|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)\\
&=n\frac{(n-k)!^2}{n!^2k!^2}n\frac{1}{n}\sum_{i_1=1}^{n}\left(\sum_{(i_2,...,i_k)\in\{1,...,n\}^{k-1}}h(\boldsymbol{i})\right)^2\\
&\quad+n\frac{(n-k)!^2}{n!^2k!^2}\frac{n(n-1)}{n^2}\left(\sum_{\boldsymbol{i}\in\{1,...,n\}^k}h(\boldsymbol{i})\right)^2\\
&=n\frac{(n-k)!^2}{n!^2k!^2}\sum_{i_1=1}^{n}\left(\sum_{(i_2,...,i_k)\in\{1,...,n\}^{k-1}}h(\boldsymbol{i})\right)^2+(n-1)\widehat{\theta}^2.
\end{aligned}$$

It follows

$$\begin{aligned}
&\mathbb{E}\left(n\left((\theta^*-\widehat{\theta})-k(\theta_1^*-\widehat{\theta})\right)^2|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)\\
=&n\mathbb{E}\left(\theta^{*2}|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)+nk^2\mathbb{E}\left(\theta_1^{*2}|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)+n(k-1)^2\widehat{\theta}^2\\
&-2kn\mathbb{E}\left(\theta^*\theta_1^*|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)+2n(k-1)\frac{n!}{n^k(n-k)!}\widehat{\theta}^2-2n(k-1)k\widehat{\theta}^2\\
=&n\widehat{\theta}^2\left(\frac{n!}{n^{2k}(n-2k)!}+k^2\frac{(n-1)}{n}+(k-1)^2+\left(2(k-1)-2k\frac{n-k}{n}\right)\frac{n!}{n^k(n-k)!}-2(k-1)k\right)\\
&+\left(\frac{k^2}{k!^2}+\frac{k^2}{k!^2}-2\frac{k^2}{k!^2}\right)\mathbb{E}(h(1,...,k)h(\mathbf{1}'))+R,
\end{aligned}$$

with $R\xrightarrow{a.s.}0$ and $\widehat{\theta}^2\xrightarrow{a.s.}\theta_0^2$. Moreover $\frac{n!}{n^{2k}(n-2k)!}=1-\frac{1}{n}\left(k(2k-1)\right)+O(n^{-2})$, $\frac{n!}{n^k(n-k)!}=1-\frac{1}{n}\left(\frac{k(k-1)}{2}\right)+O(n^{-2})$. Next

$$\begin{aligned}
&\left(\frac{n!}{n^{2k}(n-2k)!}+k^2\frac{(n-1)}{n}+(k-1)^2+\left(2(k-1)-2k\frac{n-k}{n}\right)\frac{n!}{n^k(n-k)!}-2(k-1)k\right)\\
=&\left(\frac{n!}{n^{2k}(n-2k)!}+k^2\frac{(n-1)}{n}+(k-1)^2+2\left(\frac{k^2}{n}-1\right)\frac{n!}{n^k(n-k)!}-2(k-1)k\right)\\
=&1+k^2+(k-1)^2-2-2k^2+2k+\frac{1}{n}\left(k-2k^2-k^2+2k^2+k(k-1)\right)+O(n^{-2})\\
=&O(n^{-2}).
\end{aligned}$$

We have proved that $\sqrt{n}(\theta^*-\widehat{\theta})$ converges in L^2 conditional on the data to $\sqrt{n}k(\theta_1^*-\widehat{\theta})$:

$$\mathbb{E}\left(n\left((\theta^*-\widehat{\theta})-k(\theta_1^*-\widehat{\theta})\right)^2|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)\xrightarrow{a.s.}0.$$

Characterization of the convergence in distribution for the bootstrap using the bounded-Lipschitz metric ensures that it is sufficient to prove the asymptotic normality of $\sqrt{n}k(\theta_1^*-\widehat{\theta})$. Indeed if L is a random variable whose distribution is the limit distribution of $\sqrt{n}k(\theta^*-\widehat{\theta})$ we have:

$$\begin{aligned}
&\sup_{h\in BL_1(\mathbb{R})}\left|\mathbb{E}\left(h(\sqrt{n}(\theta^*-\widehat{\theta}))|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)-\mathbb{E}(h(L))\right|\\
\leq&\sup_{h\in BL_1(\mathbb{R})}\left|\mathbb{E}\left(h(\sqrt{n}k(\theta_1^*-\widehat{\theta}))|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)-\mathbb{E}(h(L))\right|\\
&+\mathbb{E}\left(\left|\sqrt{n}((\theta^*-\widehat{\theta})-k(\theta_1^*-\widehat{\theta}))\right||(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right).
\end{aligned}$$

Equivalence of the bounded-Lipschitz and Levy criteria to metrize weak convergence entails it is sufficient to prove for every $t\in\mathbb{R}$

$$\left|\mathbb{E}\left(\exp\left(\mathfrak{i}t\sqrt{n}k(\theta_1^*-\widehat{\theta})\right)|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)-\mathbb{E}\left(\exp\left(\mathfrak{i}tL\right)\right)\right|=o_{a.s}(1),\tag{4.29}$$

to conclude. The next two substeps are devoted to proving the latter result.

Substep 2: $\lim_n\mathbb{E}\left(\left|\mathbb{E}\left(\exp\left(\mathfrak{i}t\sqrt{n}k!(\theta_1^*-\widehat{\theta})\right)|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)-e^{-t^2\mathbb{V}\mathbb{E}(h(\mathbf{1})|U_{\{1\}})/2}\right|\right)=0$**.**

Let us define

$$a_{n,i}^*=\frac{(n-k)!}{(n-1)!}\sum_{(i_2,...,i_k)\in\mathbb{I}_{n,k-1}}h(i^*,i_2,...,i_k)$$

and

$$a_{n,i}=\frac{(n-k)!}{(n-1)!}\sum_{(i_2,...,i_k)\in\mathbb{I}_{n,k-1}}h(i,i_2,...,i_k).$$

Given the sampling procedure in the bootstrap we have $\mathbb{E}(g(a^*_{n,i})|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}) = \frac{1}{n}\sum_{i=1}^n g(a_{n,i})$. Furthermore, $(a^*_{n,i})_{i=1,\dots,n}$ forms an i.i.d. sequence conditional on $(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}$. Let $Z_n = \sqrt{n}k!(\theta^*_1 - \widehat{\theta})$. Remark that Z_n can be expressed as a sum over a triangular array (reasoning conditionally on the data):

$$Z_n = \sum_{i=1}^n \frac{z_{n,i}}{\sqrt{n}} \text{ for } z_{n,i} = a^*_{n,i} - \frac{1}{n}\sum_{i'=1}^n a_{n,i'}.$$

We have for every $\epsilon > 0$ and $t \in \mathbb{R}$

$$\begin{aligned}\left|\exp\left(\frac{\mathrm{i}tz_{n,i}}{\sqrt{n}}\right) - \left(1 + \frac{\mathrm{i}tz_{n,i}}{\sqrt{n}} - \frac{t^2z^2_{n,i}}{2n}\right)\right| &\leq \min\left(\frac{|t^3z_{n,i}|^3}{\sqrt{n}^3}, \frac{t^2z^2_{n,i}}{n}\right)\\ &\leq \frac{|t^3z^3_{n,i}|}{\sqrt{n}^3}\mathbb{1}_{\{|z_{n,i}|<\epsilon\sqrt{n}\}} + \frac{t^2z^2_{n,i}}{n}\mathbb{1}_{\{|z_{n,i}|>\epsilon\sqrt{n}\}}\\ &\leq \left(\epsilon|t|^3 + t^2\mathbb{1}_{\{|z_{n,i}|>\epsilon\sqrt{n}\}}\right)\frac{z^2_{n,i}}{n}.\end{aligned}$$

Let $V_n = \mathbb{E}(z^2_{n,i}|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}) = \frac{1}{n}\sum_{i=1}^n a^2_{n,i} - \left(\frac{1}{n}\sum_{i=1}^n a_{n,i}\right)^2$ and $V = \mathbb{V}\mathbb{E}(h(\mathbf{1})|U_{\{1\}})$. Lemma S4.8 and the fact that $(h(\boldsymbol{i}))_{\boldsymbol{i}\in\mathbb{I}_k}$ is k jointly exchangeable and dissociated allow us to claim that

$$\begin{aligned}V_n &\xrightarrow{L^1,\text{a.s.}} \mathbb{E}\left[h(1,...,k)h(\mathbf{1}')\right] - \mathbb{E}\left[h(1,...,k)\right]^2\\ &= \mathbb{E}\left[\mathbb{E}\left[h(1,...,k)h(\mathbf{1}') \mid U_{\{1\}}\right]\right] - \mathbb{E}\left[\mathbb{E}\left[h(1,...,k) \mid U_{\{1\}}\right]\right]^2\\ &= \mathbb{E}\left[\mathbb{E}\left[h(1,...,k) \mid U_{\{1\}}\right]^2\right] - \mathbb{E}\left[\mathbb{E}\left[h(1,...,k) \mid U_{\{1\}}\right]\right]^2 = V,\end{aligned}$$

where the last equality can be recovered thanks to Assumption 4.1 and the almost sure representation of $(h(\boldsymbol{i}))_{\boldsymbol{i}\in\mathbb{I}_k}$.
As $\mathbb{E}(z_{n,i}|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}) = 0$, we deduce from the triangle inequality that

$$\begin{aligned}&\left|\mathbb{E}\left(\exp\left(\frac{\mathrm{i}tz_{n,i}}{\sqrt{n}}\right)\middle|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right) - \left(1 - \frac{t^2V_n}{2n}\right)\right|\\ \leq&\epsilon|t|^3\frac{V_n}{n} + \frac{t^2}{n}\mathbb{E}\left(z^2_{n,i}\mathbb{1}_{\{|z_{n,i}|>\epsilon\sqrt{n}\}}|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right),\end{aligned}$$

and then

$$\begin{aligned}&\left|\mathbb{E}\left(\exp\left(\frac{\mathrm{i}tz_{n,i}}{\sqrt{n}}\right)\middle|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right) - \left(1 - \frac{t^2V}{2n}\right)\right|\\ \leq&\epsilon|t|^3\frac{V_n}{n} + \frac{t^2}{n}\mathbb{E}\left(z^2_{n,i}\mathbb{1}_{\{|z_{n,i}|>\epsilon\sqrt{n}\}}|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right) + \frac{t^2}{2n}|V_n - V|.\end{aligned}$$

Because $|\prod_{i=1}^n a_i - \prod_{i=1}^n b_i| \leq \sum_{i=1}^n |a_i - b_i|$ if $\max_{i=1,\dots,n}\max(|a_i|,|b_i|) \leq 1$ and since the $(z_{n,i})_{i=1\dots n}$ are i.i.d. conditional on the data, we obtain

$$\begin{aligned}&\left|\mathbb{E}\left(\exp\left(\mathrm{i}tZ_n\right)|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right) - \exp\left(-\frac{t^2V}{2}\right)\right|\\ \leq&\epsilon|t|^3V_n + t^2\mathbb{E}\left(z^2_{n,1}\mathbb{1}_{\{|z_{n,1}|>\epsilon\sqrt{n}\}}|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)\\ &+ t^2|V_n - V| + \left|\exp\left(-\frac{t^2V}{2}\right) - \left(1 - \frac{t^2V}{2n}\right)^n\right|.\end{aligned}$$

A convexity argument and the Cauchy-Schwarz inequality ensure $z^2_{n,1} \leq 2a^{*2}_{n,1} + 2\left(\frac{1}{n}\sum_{i'=1}^n a_{n,i'}\right)^2 \leq$

$2a_{n,1}^{*2} + 2\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} h(\boldsymbol{i})^2$. This implies

$$\begin{aligned}
\mathbb{E}\left(z_{n,1}^2 \mathbb{1}_{\{|z_{n,1}|>\epsilon\sqrt{n}\}}\right) &\leq 2\mathbb{E}\left[\mathbb{E}\left(a_{n,1}^{*2}\mathbb{1}_{\{a_{n,1}^{*2}>\epsilon^2 n/4\}}|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)\right] \\
&+ 2\mathbb{E}\left[\mathbb{E}\left(a_{n,1}^{*2}|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)\mathbb{1}_{\left\{\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} h(\boldsymbol{i})^2>\epsilon^2 n/4\right\}}\right] \\
&+ 2\mathbb{E}\left[\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} h(\boldsymbol{i})^2\mathbb{E}\left(\mathbb{1}_{\left\{a_{n,1}^{*2}>\epsilon^2 n/4\right\}}|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}}\right)\right] \\
&+ 2\mathbb{E}\left[\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} h(\boldsymbol{i})^2\mathbb{1}_{\left\{\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} h(\boldsymbol{i})^2>\epsilon^2 n/4\right\}}\right] \\
&\leq 2\mathbb{E}\left[a_{n,1}^2\mathbb{1}_{\{a_{n,1}^2>\epsilon^2 n/4\}}\right] \\
&+ 2\mathbb{E}\left[h^2(\mathbf{1})\mathbb{1}_{\left\{\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} h^2(\boldsymbol{i})>\epsilon^2 n/4\right\}}\right] \\
&+ 2\left(1-\frac{k}{n}\right)\mathbb{E}\left[h^2(2,3,...,k+1)\mathbb{1}_{\left\{a_{n,1}^2>\epsilon^2 n/4\right\}}\right] \\
&+ 2\frac{k}{n}\mathbb{E}\left[h^2(\mathbf{1})\mathbb{1}_{\left\{a_{n,1}^2>\epsilon^2 n/4\right\}}\right] \\
&+ 2\mathbb{E}\left[h^2(\mathbf{1})\mathbb{1}_{\left\{\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} h^2(\boldsymbol{i})>\epsilon^2 n/4\right\}}\right]. \qquad (4.30)
\end{aligned}$$

Conditional on $U_{\{1\}}$, $(h(1,i_2,...,i_k))_{(i_2,...,i_k)\in\overline{(\mathbb{N}^+\setminus\{1\})^{k-1}}}$ is a jointly exchangeable and dissociated array of dimension $k-1$. Hence $a_{n,1}\xrightarrow{a.s.}\mathbb{E}\left(h(\mathbf{1})|U_{\{1\}}\right)$. Furthermore, $\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} h^2(\boldsymbol{i})\xrightarrow{a.s.}\mathbb{E}(h^2(1,...,k))$. As a result, all the indicator functions on the right-hand side of the last inequality in (4.30) tend to 0 almost surely. The dominated convergence theorem also ensures that $\mathbb{E}\left(z_{n,1}^2\mathbb{1}_{\{|z_{n,1}|>\epsilon\sqrt{n}\}}\right)\to 0$ for every $\epsilon>0$. Further, $\mathbb{E}(|V_n-V|)\to 0$ and $\left|\exp\left(-\frac{t^2V}{2}\right)-\left(1-\frac{t^2V}{2n}\right)^n\right|$ converges almost surely to 0 and is bounded. As a consequence,

$$\limsup_n \mathbb{E}\left(\left|\mathbb{E}(\exp(\mathrm{i}tZ_n)|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}})-e^{-t^2V/2}\right|\right)\leq\epsilon|t|^3.$$

Since ϵ could be chosen arbitrarily small, we finally get

$$\lim_n \mathbb{E}\left(\left|\mathbb{E}(\exp(\mathrm{i}tZ_n)|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}})-e^{-t^2V/2}\right|\right)=0.$$

Substep 3: conclusion on the almost-sure weak convergence of the bootstrap mean

We finally prove the almost-sure convergence of $\mathbb{E}(\exp(\mathrm{i}tZ_n)|(Y_{\boldsymbol{i}'})_{\boldsymbol{i}'\in\mathbb{I}_{n,k}})$, not only its convergence in L^1 as above. Recall that $V=\mathbb{V}\mathbb{E}(h(\mathbf{1})|U_{\{1\}})$ with U stemming from the AHK representation of $h(\boldsymbol{i})$. We have

$$\mathbb{E}(Z_n^2|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k})=\frac{1}{n}\sum_{i=1}^n\mathbb{V}\left(a_{n,i}^*|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)=V_n\xrightarrow{a.s.}V.$$

Given $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$, Z_n is bounded in probability: for every $\varepsilon\in(0,1)$, considering

$$\eta((Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k})=\frac{\sup_n\mathbb{E}\left(Z_n^2\middle|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)}{\varepsilon},$$

we have $\mathbb{P}\left(Z_n^2\geq\eta((Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k})|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)\leq\varepsilon$ by Markov's inequality. Given $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$, every subsequence $Z_{\sigma(n)}$ admits a further subsequence $Z_{\sigma'\circ\sigma(n)}$ that converges in distribution to $L_{\sigma'\circ\sigma}$, by Prohorov's Theorem. By Levy's criterion for weak convergence, this means that there is a set Ω' of probability one, independent of σ' and σ, such that for every $\omega\in\Omega'$, $\mathbb{E}\left(e^{itZ_{\sigma'\circ\sigma(n)}}|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}=(Y_{\boldsymbol{i}}(\omega))_{\boldsymbol{i}\in\mathbb{I}_k}\right)$ converges to $\mathbb{E}(e^{itL_{\sigma'\circ\sigma}}|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}=(Y_{\boldsymbol{i}}(\omega))_{\boldsymbol{i}\in\mathbb{I}_k})$ for every $t\in\mathbb{R}$. Note that $L_{\sigma'\circ\sigma}$ could depend on $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$. We can now

write

$$\begin{aligned}&\mathbb{E}\left[\left|\mathbb{E}[e^{itL_{\sigma'\circ\sigma}}|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}]-\exp(-t^2V/2)\right|\right]\\ \leq&\mathbb{E}\left[\left|\mathbb{E}[e^{itL_{\sigma'\circ\sigma}}|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}]-\mathbb{E}[e^{itZ_{\sigma'\circ\sigma(n)}}|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}]\right|\right]+\mathbb{E}\left[\left|\mathbb{E}[e^{itZ_{\sigma'\circ\sigma(n)}}|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}]-\exp(-t^2V/2)\right|\right].\end{aligned}$$

The first term on the right-hand side converges to 0 by dominated convergence. The second term converges to 0 by the result proved in the second substep. We finally have that almost surely, $\mathbb{E}\left[e^{itL_{\sigma'\circ\sigma}}|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right]=\exp(-t^2V/2)$ for every $t\in\mathbb{R}$, every subsequence σ and some subsequence σ'. From Urysohn's subsequence principle [see 132, Section 2.1.17, Pages 185-186], this means that almost surely, Z_n converges in distribution conditionally on $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$ to $\mathcal{N}(0,V)$. We conclude that (4.29) holds with $L\sim\mathcal{N}\left(0,\frac{k^2}{k!^2}V\right)$.

Second step: Asymptotic equicontinuity

Let $\mathcal{F}_\delta=\{f=f_1-f_2:(f_1,f_2)\in\mathcal{F}\times\mathcal{F},\mathbb{E}\left(f^2(Y_{\mathbf{1}})\right)\leq\delta^2\}$. We have to show the following almost sure convergence when $\delta\to0$

$$\limsup_{n\to\infty}\mathbb{E}\left(\sup_{\mathcal{F}_\delta}|\mathbb{G}_n^*(f)|\mid(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)\xrightarrow{a.s.}0.$$

Let $N^*=\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}$. Note that $\mathbb{E}\left[\mathbb{P}_n^*f|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right]=\mathbb{P}'_nf=\frac{1}{n^k}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}f(Y_{\boldsymbol{i}})$. By independence of the i^* with $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$, we have:

$$\begin{aligned}&\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}|\mathbb{G}_n^*f|\,\middle|\,(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right]\\ \leq&\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\sqrt{n}\left|\mathbb{P}_n^*f-\mathbb{P}'_nf\right|\,\middle|\,(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right]+\sqrt{n}\left(1-\frac{n!}{n^k(n-k)!}\right)\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}F(Y_{\boldsymbol{i}})\\ \leq&\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\sqrt{n}\left|\mathbb{P}_n^*f-\mathbb{P}'_nf\right|\,\middle|\,(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right]+\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}F(Y_{\boldsymbol{i}})\times o(1)\end{aligned}$$

Because $\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}F(Y_{\boldsymbol{i}})\xrightarrow{a.s.}\mathbb{E}\left(F(Y_{\mathbf{1}})\right)$, we only have to show that

$$\limsup_{n\to\infty}\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\sqrt{n}\left|\mathbb{P}_n^*f-\mathbb{P}'_nf\right|\,\middle|\,(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right]\xrightarrow{a.s.}0\text{ as }\delta\to0.$$

Using the symmetrization step of Lemma S4.6, we can write that for some constant C_k that depends on k only

$$\begin{aligned}&\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\sqrt{n}\left|\mathbb{P}_n^*f-\mathbb{P}'_nf\right|\,\middle|\,(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right]\\ \leq&kC_k\sqrt{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\left|\frac{1}{n}\sum_{i_1=1}^n\varepsilon_{\{i_1\}}\frac{(n-k)!}{(n-1)!}\sum_{(i_2,\ldots,i_k):\boldsymbol{i}\in\mathbb{I}_{n,k}}f\left(Y_{\boldsymbol{i}^*}\right)\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right|\,\middle|\,(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k},N^*>0\right]\mathbb{P}\left(N^*>0\right).\end{aligned}$$

We have

$$\begin{aligned}&\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\left|\frac{1}{n}\sum_{i_1=1}^n\varepsilon_{\{i_1\}}\frac{(n-k)!}{(n-1)!}\sum_{(i_2,\ldots,i_k):\boldsymbol{i}\in\mathbb{I}_{n,k}}f\left(Y_{\boldsymbol{i}^*}\right)\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right|\,\middle|\,(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k},(\boldsymbol{i}^*)_{\boldsymbol{i}\in\mathbb{I}_{n,k}},N^*>0\right]\\ \leq&\frac{4\sqrt{2}}{\sqrt{n}}\int_0^{\sigma_{1,2}}\sqrt{\log2N(\varepsilon,\mathcal{F}_\delta,||.||_{1,2})}d\varepsilon,\end{aligned}$$

for $||f||_{1,2}^2=\frac{1}{n}\sum_{i_1=1}^n\left(\frac{(n-k)!}{(n-1)!}\sum_{(i_2,\ldots,i_k):\boldsymbol{i}\in\mathbb{I}_{n,k}}f(Y_{\boldsymbol{i}^*})\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right)^2$ and $\sigma_{1,2}^2=\sup_{\mathcal{F}_\delta}||f||_{1,2}^2$. We now reason conditional on $N^*>0$. The Cauchy-Schwarz inequality ensures $||f||_{1,2}^2\leq N^*||f||_{\mathbb{P}_n^*,2}^2$ for $||f||_{\mathbb{P}_n^*,2}^2=$

$N^{*-1}\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} f^2(Y_{\boldsymbol{i}^*})\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}$. It follows that (see Point 1 of Lemma S4.11)

$$\sigma_{1,2}^2 \leq \sigma_n^{*2} = \sup_{\mathcal{F}_\delta} N^* ||f||^2_{\mathbb{P}_n^*,2},$$

$$\text{and } N(\varepsilon, \mathcal{F}_\delta, ||.||_{1,2}) \leq N(\varepsilon, \mathcal{F}_\delta, N^{*1/2}||.||_{\mathbb{P}_n^*,2}) \leq N(\varepsilon N^{*-1/2}, \mathcal{F}_\delta, ||.||_{\mathbb{P}_n^*,2}).$$

Monotonicity of the integral, Points 3 and 4 of Lemma S4.11 and the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ entail

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta} |\mathbb{G}_n^* f| \,\middle|\, (Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right] \\
\leq& K_k' \mathbb{E}\left[\sigma_n^* + \int_0^{\sigma_n^*} \sqrt{\log N\left(4\varepsilon N^{*-1/2}, \mathcal{F}, ||.||_{\mathbb{P}_n^*,2}\right)} d\varepsilon \,\middle|\, (Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}, N^* > 0\right] \mathbb{P}\left(N^* > 0\right),
\end{aligned}$$

for some constant K_k' depending only on k. Furthermore, when $N^* > 0$ the following holds:

$$\begin{aligned}
&\int_0^{\sigma_n^*} \sqrt{\log N\left(4\varepsilon N^{*-1/2}, \mathcal{F}, ||.||_{\mathbb{P}_n^*,2}\right)} d\varepsilon \\
=& \int_0^{\sigma_n^*} \sqrt{\log N\left(\varepsilon ||F||_{\mathbb{P}_n^*,2}/(4N^{*1/2}||F||_{\mathbb{P}_n^*,2}), \mathcal{F}, ||.||_{\mathbb{P}_n^*,2}\right)} d\varepsilon \\
=& 4N^{*1/2}||F||_{\mathbb{P}_n^*,2} \int_0^{\sigma_n^*/(4N^{*1/2}||F||_{\mathbb{P}_n^*,2})} \sqrt{\log N\left(\varepsilon ||F||_{\mathbb{P}_n^*,2}, \mathcal{F}, ||.||_{\mathbb{P}_n^*,2}\right)} d\varepsilon \\
\leq& 4\sqrt{N^*||F||^2_{\mathbb{P}_n^*,2}} J_{\mathcal{F}}\left(\frac{\sqrt{\sigma_n^{*2}}}{4\sqrt{N^*||F||^2_{\mathbb{P}_n^*,2}}}\right).
\end{aligned}$$

This, Lemma S4.10, the fact that $\mathbb{E}\left(\sigma_n^{*2}|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}, N^* > 0\right) = \mathbb{E}\left(\sigma_n^{*2}|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)/\mathbb{P}\left(N^* > 0\right)$ and $\mathbb{E}\left(N^*||F||^2_{\mathbb{P}_n^*,2}|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}, N^* > 0\right) = \frac{1}{n^k}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} F^2(Y_{\boldsymbol{i}})/\mathbb{P}\left(N^* > 0\right)$ and Jensen's inequality thus ensure

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta} |\mathbb{G}_n^* f| \,\middle|\, (Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right] \\
\leq& K_k' \left(\mathbb{E}\left(\sigma_n^{*2}|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)^{1/2} + \left(\frac{1}{n^k}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} F^2(Y_{\boldsymbol{i}})\right)^{1/2} J_{\mathcal{F}}\left(\frac{\mathbb{E}\left(\sigma_n^{*2}|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)^{1/2}}{4\left(\frac{1}{n^k}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} F^2(Y_{\boldsymbol{i}})\right)^{1/2}}\right)\sqrt{\mathbb{P}\left(N^* > 0\right)}\right).
\end{aligned}$$

Since $\frac{1}{n^k}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} F^2(Y_{\boldsymbol{i}}) \xrightarrow{a.s.} \mathbb{E}\left(F^2(Y_{\boldsymbol{1}})\right)$, we only have to show that

$$\limsup_{n\to\infty} \mathbb{E}\left(\sigma_n^{*2}|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right) \xrightarrow{a.s.} 0 \text{ as } \delta \downarrow 0.$$

We have

$$\begin{aligned}
\sigma_n^{*2} = \sup_{\mathcal{F}_\delta} |\mathbb{P}_n^* f^2| &\leq \sup_{\mathcal{F}_\delta} \left|\mathbb{P}_n^* f^2 - \mathbb{P}_n f^2\right| + \sup_{\mathcal{F}_\delta} \left|\mathbb{P}_n f^2 - P f^2\right| + \delta^2 \\
&\leq \sup_{\mathcal{F}_\infty} \left|\mathbb{P}_n^* f^2 - \mathbb{P}_n f^2\right| + \sup_{\mathcal{F}_\infty} \left|\mathbb{P}_n f^2 - P f^2\right| + \delta^2.
\end{aligned}$$

Point 5 of Lemma S4.11 entails

$$\sup_Q N\left(\eta||4F^2||_{Q,1}, \mathcal{F}_\infty^2, ||.||_{Q,1}\right) < +\infty \text{ for every } \eta > 0.$$

Theorem 4.1 and Lemma S4.6 imply

$$\mathbb{E}\left(\sup_{\mathcal{F}_\infty} \left|\mathbb{P}_n^* f^2 - \mathbb{P}_n f^2\right| \mid (Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right) \xrightarrow{a.s.} 0 \quad \text{and} \quad \sup_{\mathcal{F}_\infty} \left|\mathbb{P}_n f^2 - P f^2\right| \xrightarrow{a.s.} 0,$$

which finally leads to

$$\lim_{\delta\downarrow 0} \limsup_{n\to+\infty} \mathbb{E}\left(\sigma_n^{*2} \mid (Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right) = 0 \quad \text{a.s.}$$

4.7.1.4 Theorem 4.3

The proof is the same as that of Theorem 13.4 in [102], with one change only: we have to check that $\mathbb{G}$, the limit of $\theta \mapsto \sqrt{n}(\Psi_n(\theta) - \Psi(\theta))$, is continuous. Given the kernel of $\mathbb{G}$, it suffices to check that for all $(\pi, \pi') \in \mathfrak{S}(\{\mathbf{1}\}) \times \mathfrak{S}(\{\mathbf{1}'\})$,

$$\sup_{h \in \mathcal{H}} \left| \mathbb{C}ov \left([\psi_{\theta,h} - \psi_{\theta_0,h}](Y_{\pi(\mathbf{1})}), [\psi_{\theta,h} - \psi_{\theta_0,h}](Y_{\pi'(\mathbf{1}')}) \right) \right| \to 0. \tag{4.31}$$

By Cauchy-Schwarz's inequality and joint exchangeability, this covariance is smaller than

$$\mathbb{E}\left\{ [\psi_{\theta,h} - \psi_{\theta_0,h}]^2 (Y_{\pi(\mathbf{1})}) \right\} = P(\psi_{\theta,h} - \psi_{\theta_0,h})^2.$$

Therefore, Condition 4 ensures that (4.31) holds. The result follows.

4.7.1.5 Theorem 4.4

The first result follows by Theorem 4.1.2 because the class $\{u \mapsto 1\{u \leq y\} : y \in \mathbb{R}^p\}$ is pointwise measurable and satisfies Assumption 4.4. The second point follows directly from Point 1 and the functional delta method, see e.g. Theorem 20.8 in [136]. Finally, Point 3 follows from Theorem 4.2 and the functional delta method for the bootstrap, see e.g. Theorem 23.9 in [136].

4.7.2 Proofs of the extensions

4.7.2.1 Theorem 4.5

4.7.2.1.1 Uniform law of large numbers

We remark that $\sup_{f \in \mathcal{F}} |\widetilde{\mathbb{P}}_n f - \widetilde{P} f| = \sup_{\widetilde{f} \in \widetilde{\mathcal{F}}} |\mathbb{P}_n \widetilde{f} - P\widetilde{f}|$. Following the same reasoning as in the proof of Theorem 4.1, for every positive M and η_1 (with η_1 possibly random) and some constants $K_{r,k}$, there exists a jointly exchangeable and dissociated array $(\widetilde{Y}^r_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{I}_k} = (N^r_{\boldsymbol{i}}, (Y^r_{\boldsymbol{i},\ell})_{\ell \geq 1})_{\boldsymbol{i} \in \mathbb{I}_k}$ such that $\widetilde{Y}^r_{\boldsymbol{i}} \overset{d}{=} (N_{\boldsymbol{i}}, (Y_{\boldsymbol{i},\ell})_{\ell \geq 1})_{\boldsymbol{i} \in \mathbb{I}_k}$ for all $\boldsymbol{i} \in \mathbb{I}_{n,k}$ and

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\widetilde{f} \in \widetilde{\mathcal{F}}} |\mathbb{P}_n \widetilde{f} - P\widetilde{f}| \right] \\
\leq &\mathbb{E}\left[\widetilde{F}(\widetilde{Y}_{\mathbf{1}}) \mathbb{1}_{\{\widetilde{F}(\widetilde{Y}_{\mathbf{1}}) > M\}} \right] \\
&+ \sum_{r=1}^{k} \sum_{\boldsymbol{e} \in \mathcal{E}_r} K_{r,k} \mathbb{E}\left[\sup_{\mathcal{F}} \left| \frac{(n-k)!}{n!} \sum_{\boldsymbol{i} \in \mathbb{I}_{n,k}} \varepsilon_{\{\boldsymbol{i} \odot \boldsymbol{e}\}^+} \widetilde{f}(\widetilde{Y}^r_{\boldsymbol{i}}) \mathbb{1}_{\{\widetilde{F}(\widetilde{Y}^r_{\boldsymbol{i}}) \leq M\}} \right| \middle| \overline{N^r_1} > 0 \right] \mathbb{P}\left(\overline{N^r_1} > 0 \right) \\
\leq &\mathbb{E}\left[\widetilde{F}(\widetilde{Y}_{\mathbf{1}}) \mathbb{1}_{\{\widetilde{F}(\widetilde{Y}_{\mathbf{1}}) > M\}} \right] \\
&+ \sum_{r=1}^{k} \sum_{\boldsymbol{e} \in \mathcal{E}_r} K_{r,k} \mathbb{E}\left[\sqrt{2 \log 2N(\eta_1, \widetilde{\mathcal{F}}, ||.||_{\boldsymbol{e},M,1})} M \frac{\sqrt{(n-r)!r!}}{\sqrt{n!}} + \eta_1 \middle| \overline{N^r_1} > 0 \right] \mathbb{P}\left(\overline{N^r_1} > 0 \right),
\end{aligned}$$

where $\overline{N^r_p} = \frac{(n-k)!}{n!} \sum_{\boldsymbol{i} \in \mathbb{I}_{n,k}} (N^r_{\boldsymbol{i}})^p$. Moreover, $||\widetilde{f}||_{\boldsymbol{e},M,1} \leq \overline{N^r_1} ||f||_{\mathbb{Q}^r_n,1}$ with

$$\mathbb{Q}^r_n = \frac{1}{\sum_{\boldsymbol{i} \in \mathbb{I}_{n,k}} (N^r_{\boldsymbol{i}})^2} \sum_{\boldsymbol{i} \in \mathbb{I}_{n,k}} N^r_{\boldsymbol{i}} \sum_{\ell=1}^{N^r_{\boldsymbol{i}}} \delta_{Y^r_{\boldsymbol{i},\ell}}.$$

Letting $\eta_1 = \eta \overline{N^r_1} ||F||_{\mathbb{Q}^r_n,1}$ for an arbitrary $\eta > 0$, we have $N(\eta_1, \widetilde{\mathcal{F}}, ||.||_{\boldsymbol{e},M,1}) \leq N(\eta_1, \mathcal{F}, \overline{N^r_1} ||.||_{\mathbb{Q}^r_n,1}) = N(\overline{N^r_1}^{-1} \eta_1, \mathcal{F}, ||.||_{\mathbb{Q}^r_n,1})$ whenever $\overline{N^r_1} > 0$.

Combining this insight with the fact that $\mathbb{E}\left[||\widetilde{F}||_{\mathbb{Q}_n^r,1} \mid \overline{N_1^r} > 0\right] = \mathbb{E}\left[\widetilde{F}(\widetilde{Y}_{\mathbf{1}})\right] / \mathbb{P}\left(\overline{N_1^r} > 0\right)$, we get

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\widetilde{\mathcal{F}}}\left|\mathbb{P}_n\widetilde{f} - P\widetilde{f}\right|\right] \\
\leq &\mathbb{E}\left[\widetilde{F}(\widetilde{Y}_{\mathbf{1}})\mathbb{1}_{\left\{\widetilde{F}(\widetilde{Y}_{\mathbf{1}})>M\right\}}\right] + \sum_{r=1}^{k}\sum_{\boldsymbol{e}\in\mathcal{E}_r} K_{r,k}\sqrt{2\log 2 \sup_Q N(\eta||F||_{Q,1},\mathcal{F},||.||_{Q,1})}M\frac{\sqrt{(n-r)!r!}}{\sqrt{n!}} \\
&+ \eta\sum_{r=1}^{k}\sum_{\boldsymbol{e}\in\mathcal{E}_r} K_{r,k}\mathbb{E}\left[\widetilde{F}(\widetilde{Y}_{\mathbf{1}})\right].
\end{aligned}$$

Considering M sufficiently large and η sufficiently small and next n tending to ∞ we deduce that $\mathbb{E}\left[\sup_{\widetilde{\mathcal{F}}}|\mathbb{P}_n\widetilde{f} - P\widetilde{f}|\right]$ tends to 0 as $n \to \infty$.

Let Σ_n be the σ-algebra generated by $\mathcal{H}_n$ the set of functions g from $\mathcal{D}^{\mathbb{I}_k}$ to $\mathbb{R}$ that are invariant by the action of any permutation π on $\mathbb{N}^+$ such that $\pi(j) = j$ for $j \geq n$:

$$g\left((\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right) = g\left((\widetilde{Y}_{\pi(\boldsymbol{i})})_{\boldsymbol{i}\in\mathbb{I}_k}\right).$$

Following the same reasoning as in the proof of Theorem 4.1, we conclude that $\left(\sup_{\widetilde{\mathcal{F}}}|\mathbb{P}_n\widetilde{f} - P\widetilde{f}|, \Sigma_n\right)_{n\geq 1}$ is a backwards submartingale ensuring the almost sure convergence of $\sup_{\widetilde{\mathcal{F}}}|\mathbb{P}_n\widetilde{f} - P\widetilde{f}|$.

4.7.2.1.2 Uniform central limit theorem

The pointwise weak convergence is ensured by the first step of the proof of Theorem 4.1.2 applied to the class $\widetilde{\mathcal{F}}$ because for every $f \in \mathcal{F}$ we have $\mathbb{E}\left[\left(\sum_{\ell=1}^{N_{\mathbf{1}}} f(Y_{\mathbf{1},\ell})\right)^2\right] < +\infty$. We just have to show the asymptotic equicontinuity and total boundedness of $\widetilde{\mathcal{F}}$.

Reasoning as in the proof of Theorem 4.1, we get

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\left|\widetilde{\mathbb{G}}_n f\right|\right] = \mathbb{E}\left[\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\delta}\left|\mathbb{G}_n\widetilde{f}\right|\right] = \sum_{r=1}^{k} O\left(\mathbb{E}\left(\int_0^{\widetilde{\sigma}_n^r}\sqrt{\log 2N\left(\varepsilon,\widetilde{\mathcal{F}}_\delta,||.||_{\mu_n^r,2}\right)}d\varepsilon\right)\right),$$

with $\mu_n^r = \frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\delta_{\{(N_{\boldsymbol{i}}^r,(Y_{\boldsymbol{i},\ell}^r)_{N_{\boldsymbol{i}}^r\geq\ell\geq 1})\}}$ and $(\widetilde{\sigma}_n^r)^2 = \sup_{\widetilde{\mathcal{F}}_\delta}||\widetilde{f}||_{\mu_n^r,2}^2$. If $\overline{N_2^r} = 0$, we remark that $\int_0^{\widetilde{\sigma}_n^r}\sqrt{\log 2N\left(\varepsilon,\widetilde{\mathcal{F}}_\delta,||.||_{\mu_n^r,2}\right)}d\varepsilon = 0$. As a result, we can write

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\left|\widetilde{\mathbb{G}}_n f\right|\right] = \sum_{r=1}^{k} O\left(\mathbb{E}\left(\int_0^{\widetilde{\sigma}_n^r}\sqrt{\log 2N\left(\varepsilon,\widetilde{\mathcal{F}}_\delta,||.||_{\mu_n^r,2}\right)}d\varepsilon\middle|\overline{N_2^r} > 0\right)\mathbb{P}\left(\overline{N_2^r} > 0\right)\right).$$

Reasoning conditional on $\overline{N_2^r} > 0$, we let $\mathbb{Q}_n^r = \frac{1}{\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}(N_{\boldsymbol{i}}^r)^2}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} N_{\boldsymbol{i}}^r\sum_{\ell=1}^{N_{\boldsymbol{i}}^r}\delta_{\{Y_{\boldsymbol{i},\ell}^r\}}$. For every $f \in \mathcal{F}_\delta$ and $\widetilde{f}$ the corresponding element in $\widetilde{\mathcal{F}}_\delta$, we have by the Cauchy-Schwarz inequality

$$||\widetilde{f}||_{\mu_n^r,2}^2 \leq \overline{N_2^r}||f||_{\mathbb{Q}_n^r,2}^2,$$

and next $N(\varepsilon,\widetilde{\mathcal{F}}_\delta,||.||_{\mu_n^r,2}) \leq N(\varepsilon,\mathcal{F}_\delta,\overline{N_2^r}^{1/2}||.||_{\mathbb{Q}_n^r,2})$. Moreover, Points 1, 3 and 4 of Lemma S4.11 ensure that $N(\varepsilon,\widetilde{\mathcal{F}}_\delta,||.||_{\mu_n^r,2}) \leq N^2(\varepsilon/4\overline{N_2^r}^{1/2},\mathcal{F},||.||_{\mathbb{Q}_n^r,2})$. The inequality $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$, Lemma S4.10, the fact that $\mathbb{E}\left[\widetilde{\sigma}_n^r \mid \overline{N_2^r} > 0\right] = \mathbb{E}\left[\widetilde{\sigma}_n^r\right] / \mathbb{P}\left(\overline{N_2^r} > 0\right)$, $\mathbb{E}\left[N_{\mathbf{1}}^r\sum_{\ell=1}^{N_{\mathbf{1}}^r}F^2(Y_{\mathbf{1},\ell}^r) \mid \overline{N_2^r} > 0\right]$
$= \mathbb{E}\left[N_{\mathbf{1}}\sum_{\ell=1}^{N_{\mathbf{1}}}F^2(Y_{\mathbf{1},\ell})\right] / \mathbb{P}\left(\overline{N_2^r} > 0\right)$ and Jensen's inequality imply

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\left|\widetilde{\mathbb{G}}_n f\right|\right] \leq \sum_{r=1}^{k} O\left(\mathbb{E}\left[(\widetilde{\sigma}_n^r)^2\right]^{1/2} + \mathbb{E}\left[N_{\mathbf{1}}\sum_{\ell=1}^{N_{\mathbf{1}}}F^2(Y_{\mathbf{1},\ell})\right]^{1/2} J_{\mathcal{F}}\left(\frac{\mathbb{E}\left[(\widetilde{\sigma}_n^r)^2\right]^{1/2}}{4\mathbb{E}\left(N_{\mathbf{1}}\sum_{\ell=1}^{N_{\mathbf{1}}}F^2(Y_{\mathbf{1},\ell})\right)^{1/2}}\right)\right).$$

To prove asymptotic equicontinuity, it is sufficient to show that $\lim_{\delta\downarrow 0}\limsup_{n\to\infty}\mathbb{E}\left[(\widetilde{\sigma}_n^r)^2\right]=0$ for every $r=1,...,k$. We have:

$$\mathbb{E}\left[(\widetilde{\sigma}_n^r)^2\right]\leq\mathbb{E}\left[\sup_{f\in\widetilde{\mathcal{F}}_\delta}\left|\mu_n^r f^2-Pf^2\right|\right]+\delta^2\leq\mathbb{E}\left[\sup_{f\in\widetilde{\mathcal{F}}_\infty}\left|\mu_n^r f^2-Pf^2\right|\right]+\delta^2.$$

For $r'=1,...,k$, we define $(\widetilde{Y}_{\boldsymbol{i}}^{r,r'})_{\boldsymbol{i}\in\mathbb{I}_k}=(N_{\boldsymbol{i}}^{r,r'},(Y_{\boldsymbol{i},\ell}^{r,r'})_{\ell\geq 1})_{\boldsymbol{i}\in\mathbb{I}_k}$ to be a jointly exchangeable and dissociated array such that $\widetilde{Y}_{\boldsymbol{i}}^{r,r'}\overset{d}{=}(N_{\boldsymbol{i}},(Y_{\boldsymbol{i},\ell})_{\ell\geq 1})_{\boldsymbol{i}\in\mathbb{I}_k}$ for all $\boldsymbol{i}\in\mathbb{I}_{n,k}$. We finally let $\overline{N_2^{r,r'}}=\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}(N_{\boldsymbol{i}}^{r,r'})^2$. Following the same reasoning as in the proof of Theorem 4.1, we have, for every positive M and η_1 (with η_1 possibly random) and some constants $K_{r,r',k}$,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\widetilde{\mathcal{F}}_\infty}|\mu_n^r f^2-Pf^2|\right]\\
\leq&\mathbb{E}\left[\left(\widetilde{F}(\widetilde{Y}_{\mathbf{1}})\right)^2\mathbb{1}_{\left\{\left(\widetilde{F}(\widetilde{Y}_{\mathbf{1}})\right)^2>M\right\}}\right]\\
&+\sum_{r'=1}^{k}\sum_{\boldsymbol{e}\in\mathcal{E}_{r'}}K_{r,r',k}\mathbb{E}\left[\sup_{f\in\mathcal{F}_\infty}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}\left(\widetilde{f}\left(\widetilde{Y}_{\boldsymbol{i}}^{r,r'}\right)\right)^2\mathbb{1}_{\left\{\left(\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}}^{r,r'})\right)^2\leq M\right\}}\right|\middle|\overline{N_2^{r,r'}}>0\right]\\
&\times\mathbb{P}\left(\overline{N_2^{r,r'}}>0\right)\\
\leq&\mathbb{E}\left[\left(\widetilde{F}(\widetilde{Y}_{\mathbf{1}})\right)^2\mathbb{1}_{\left\{\left(\widetilde{F}(\widetilde{Y}_{\mathbf{1}})\right)^2>M\right\}}\right]\\
&+\sum_{r'=1}^{k}\sum_{\boldsymbol{e}\in\mathcal{E}_r}K_{r,r',k}\mathbb{E}\left[\sqrt{2\log 2N(\eta_1,\widetilde{\mathcal{F}}_\infty^2,||.||_{\boldsymbol{e},M,1})}M\frac{\sqrt{(n-r')!r'!}}{\sqrt{n!}}+\eta_1\middle|\overline{N_2^{r,r'}}>0\right]\mathbb{P}\left(\overline{N_2^{r,r'}}>0\right),
\end{aligned}$$

with $\widetilde{\mathcal{F}}_\infty^2=\left\{g:g(n,y_1,...,y_n)=\left[\sum_{\ell=1}^n f(y_\ell)\right]^2,f\in\mathcal{F}_\infty\right\}$ and the seminorm $||\cdot||_{\boldsymbol{e},M,1}$ defined by

$$||g||_{\boldsymbol{e},M,1}=\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\overrightarrow{\mathbb{I}_{n,r'}}}\left|\sum_{\pi\in\mathfrak{S}_{r'}}\sum_{\boldsymbol{i}'\in\overline{(\{1,...,n\}\setminus\{\boldsymbol{i}\})^{k-r'}}}\left[\widetilde{f}\left(\widetilde{Y}^{r,r'}_{(\boldsymbol{i}_\pi)^{\boldsymbol{e}}+\boldsymbol{i}'^{(1-\boldsymbol{e})}}\right)\right]^2\mathbb{1}_{\left\{\left[\widetilde{F}\left(\widetilde{Y}^{r,r'}_{(\boldsymbol{i}_\pi)^{\boldsymbol{e}}+\boldsymbol{i}'^{(1-\boldsymbol{e})}}\right)\right]^2\leq M\right\}}\right|$$

for $g\in\widetilde{\mathcal{F}}_\infty^2$ and its corresponding $f\in\mathcal{F}_\infty$. When $\overline{N_2^{r,r'}}>0$, we have $||g||_{\boldsymbol{e},M,1}\leq\overline{N_2^{r,r'}}||f^2||_{\mathbb{Q}_n^{r,r'},1}$, $||f^2||_{\mathbb{Q}_n^{r,r'},1}=\frac{1}{\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}(N^{r,r'})_{\boldsymbol{i}}^2}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}N_{\boldsymbol{i}}^{r,r'}\sum_{\ell=1}^{N_{\boldsymbol{i}}^{r,r'}}f^2(Y_{\boldsymbol{i},\ell}^{r,r'})$ and

$$N(\eta_1,\widetilde{\mathcal{F}}_\infty^2,||.||_{\boldsymbol{e},M,1})\leq N(\eta_1,\mathcal{F}_\infty^2,\overline{N_2}||.||_{\mathbb{Q}_n^{r,r'},1}).$$

Let $\eta_1=8\eta\overline{N_2^{r,r'}}||F^2||_{\mathbb{Q}_n^{r,r'},1}$ for an arbitrary $\eta>0$. Point 4 of Lemma S4.11 ensures $N(\eta_1,\widetilde{\mathcal{F}}_\infty^2,||.||_{\boldsymbol{e},M,1})\leq N^2(\eta||F||_{\mathbb{Q}_n^{r,r'},2},\mathcal{F},||.||_{\mathbb{Q}_n^{r,r'},2})$. Combining this insight with the fact that

$$\mathbb{E}\left[\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}N_{\boldsymbol{i}}^{r,r'}\sum_{\ell=1}^{N_{\boldsymbol{i}}^{r,r'}}F^2(Y_{\boldsymbol{i},\ell}^{r,r'})\mid\overline{N_2^{r,r'}}>0\right]\mathbb{P}\left(\overline{N_2^{r,r'}}>0\right)=\mathbb{E}\left[N_{\mathbf{1}}\sum_{\ell=1}^{N_{\mathbf{1}}}F^2(Y_{\mathbf{1},\ell})\right],$$

we get

$$\begin{aligned}
\mathbb{E}\left[\sup_{\widetilde{\mathcal{F}}_\infty}\left|\mu_n^r f^2-Pf^2\right|\right]\leq&4\mathbb{E}\left[\left(\widetilde{F}(\widetilde{Y}_{\mathbf{1}})\right)^2\mathbb{1}_{\left\{\left(\widetilde{F}(\widetilde{Y}_{\mathbf{1}})\right)^2>M\right\}}\right]\\
&+4\sum_{r'=1}^{k}\sum_{\boldsymbol{e}\in\mathcal{E}_r}K_{r,r',k}\sqrt{2\log 2\sup_Q N^2(\eta||F||_{Q,2},\mathcal{F},||.||_{Q,2})}M\frac{\sqrt{(n-r')!r'!}}{\sqrt{n!}}\\
&+8\eta\sum_{r'=1}^{k}\sum_{\boldsymbol{e}\in\mathcal{E}_r}K_{r,r',k}\mathbb{E}\left[N_{\mathbf{1}}\sum_{\ell=1}^{N_{\mathbf{1}}}F^2(Y_{\mathbf{1},\ell})\right].
\end{aligned}$$

Note that $\sup_Q N^2(\eta||F||_{Q,2},\mathcal{F},||.||_{Q,2}) = \left(\sup_Q N(\eta||F||_{Q,2},\mathcal{F},||.||_{Q,2})\right)^2 < +\infty$ for every η. Considering M sufficiently large and η sufficiently small and next n tending to ∞ we deduce that $\mathbb{E}\left[\sup_{\widetilde{\mathcal{F}}_\infty}\left|\mu_n^r f^2 - Pf^2\right|\right]$ tends to 0 as $n\to\infty$ for every $r = 1,...,k$.

To conclude the proof of weak convergence, we have to verify total boundedness. By the Markov inequality, we have just shown $\sup_{\widetilde{\mathcal{F}}_\infty}\left|\mu_n^r f^2 - Pf^2\right| = o_p(1)$ for $r = 1,...,k$. Fixing r, this entails that for every $\varepsilon > 0$ there exists $R_\varepsilon = o_p(1)$ such that for every pair $(f_1,f_2)\in\mathcal{F}\times\mathcal{F}$

$$\mathbb{E}\left[\left(\widetilde{f}_1(\widetilde{Y}_{\mathbf{1}}) - \widetilde{f}_2(\widetilde{Y}_{\mathbf{1}})\right)^2\right] \leq ||\widetilde{f}_1 - \widetilde{f}_2||^2_{\mu_n^r,2} + R_\varepsilon.$$

For every $c > 1$, by definition of covering numbers

$$N(c\varepsilon,\widetilde{\mathcal{F}},||.||_{P,2}) \leq N\left(\varepsilon,\widetilde{\mathcal{F}},||.||_{\mu_n^r,2}\right) + o_p(1).$$

If $\overline{N_2^r}||F||^2_{\mathbb{Q}_n^r,2} > 0$, let $U = \varepsilon/(2\overline{N_2^r}^{1/2}||F||_{\mathbb{Q}_n^r,2})$. We have $\overline{N_2^r}||F||^2_{\mathbb{Q}_n^r,2} \xrightarrow{a.s.} \mathbb{E}\left(N_{\mathbf{1}}\sum_{\ell=1}^{N_{\mathbf{1}}} F^2(Y_{\mathbf{1},\ell})\right) > 0$ by the almost sure convergence of the mean of jointly exchangeable arrays [66] and ergodicity for dissociated arrays [96]. Starting from the last inequality, we obtain for every $\varepsilon > 0$

$$\begin{aligned}
N(\varepsilon,\widetilde{\mathcal{F}},||.||_{P,2}) &\leq N\left(\frac{\varepsilon}{2},\widetilde{\mathcal{F}},||.||_{\mu_n^r,2}\right) + o_p(1)\\
&\leq N\left(\frac{\varepsilon}{2},\mathcal{F},\overline{N_2^r}^{1/2}||.||_{\mathbb{Q}_n^r,2}\right) + o_p(1)\\
&= N\left(U||F||_{\mathbb{Q}_n^r,2},\mathcal{F},||.||_{\mathbb{Q}_n^r,2}\right)\mathbb{1}_{\{\overline{N_2^r}||F||^2_{\mathbb{Q}_n^r,2}>0\}} + \mathbb{1}_{\{\overline{N_2^r}||F||^2_{\mathbb{Q}_n^r,2}=0\}} + o_p(1)\\
&\leq \sup_Q N\left(U||F||_{Q,2},\mathcal{F},||.||_{Q,2}\right)\mathbb{1}_{\{\overline{N_2^r}||F||^2_{\mathbb{Q}_n^r,2}>0\}} + o_p(1)\\
&< +\infty,
\end{aligned}$$

where the second inequality is a consequence of the Cauchy-Schwarz inequality and the equality on the third line is a consequence of Point 1 of Lemma S4.11. Hence, total boundedness holds.

4.7.2.2 Convergence of the bootstrap process

The triangle inequality ensures that for every $f\in\mathcal{F}$, we have $\mathbb{E}\left(\left(\widetilde{f}(\widetilde{Y}_{\mathbf{1}})\right)^2\right) \leq \mathbb{E}\left(\left(\widetilde{F}(\widetilde{Y}_{\mathbf{1}})\right)^2\right) < +\infty$. The pointwise weak convergence thus follows from Theorem 4.2 applied to a finite class. The total boundedness of $(\widetilde{\mathcal{F}},||.||_{P,2})$ has already been proved (see the proof of Theorem 4.5.2). As a result, to prove weak convergence we only have to prove asymptotic equicontinuity.

Applying the symmetrization argument used in the proof of Lemma S4.6, we have the following inequality for some number C_k depending only on k

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\left|\widetilde{\mathbb{G}}_n^* f\right|\,\middle|\,(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right]\\
\leq& kC_k\sqrt{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\left|\frac{1}{n}\sum_{i_1=1}^n \varepsilon_{\{i_1\}}\frac{(n-k)!}{(n-1)!}\sum_{(i_2,...,i_k):\boldsymbol{i}\in\mathbb{I}_{n,k}}\widetilde{f}(\widetilde{Y}_{\boldsymbol{i}^*})\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right|\,\middle|\,(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right].
\end{aligned}$$

If $\overline{N_2^*} = \frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} N_{\boldsymbol{i}^*}^2\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}$ is null, $\widetilde{f}(\widetilde{Y}_{\boldsymbol{i}^*})$ is null for every $\boldsymbol{i}\in\mathbb{I}_{n,k}$ and

$$\sup_{f\in\mathcal{F}_\delta}\left|\frac{1}{n}\sum_{i_1=1}^n \varepsilon_{\{i_1\}}\frac{(n-k)!}{(n-1)!}\sum_{(i_2,...,i_k)\in\mathbb{I}_{n,k-1}}\widetilde{f}(\widetilde{Y}_{\boldsymbol{i}^*})\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right| = 0.$$

Otherwise $\overline{N_2^*} > 0$ and

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\left|\frac{1}{n}\sum_{i_1=1}^{n}\varepsilon_{\{i_1\}}\frac{(n-k)!}{(n-1)!}\sum_{(i_2,\ldots,i_k)\in\mathbb{I}_{n,k-1}}\widetilde{f}(\widetilde{Y}_{\boldsymbol{i}^*})\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right|\middle|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k},(\boldsymbol{i}^*)_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\right]$$
$$\leq\frac{4\sqrt{2}}{\sqrt{n}}\int_0^{\sigma_{1,2}^*}\sqrt{\log 2N(\varepsilon,\widetilde{\mathcal{F}}_\delta,||.||_{1,2})}d\varepsilon,$$

for $||\widetilde{f}||_{1,2}^2 = \frac{1}{n}\sum_{i_1=1}^{n}\left(\frac{(n-k)!}{(n-1)!}\sum_{(i_2,\ldots,i_k):\boldsymbol{i}\in\mathbb{I}_{n,k}}\widetilde{f}(\widetilde{Y}_{\boldsymbol{i}^*})\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right)^2$ and $\sigma_{1,2}^{*2} = \sup_{\widetilde{\mathcal{F}}_\delta}||\widetilde{f}||_{1,2}^2$. The Cauchy-Schwarz inequality ensures that for every $f\in\mathcal{F}_\delta$, $||\widetilde{f}||_{1,2}^2 \leq \overline{N_2^*}||f||_{\mathbb{Q}_n^*,2}^2$ and $||f||_{\mathbb{Q}_n^*,2}^2 = \frac{1}{\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}N_{\boldsymbol{i}^*}^2}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}N_{\boldsymbol{i}^*}\sum_{\ell=1}^{N_{\boldsymbol{i}^*}}f^2(Y_{\boldsymbol{i}^*,\ell})\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}$.

It follows from Point 1 of Lemma S4.11 that

$$N(\varepsilon,\widetilde{\mathcal{F}}_\delta,||.||_{1,2}) \leq N(\varepsilon,\mathcal{F}_\delta,\overline{N_2^*}^{1/2}||.||_{\mathbb{Q}_n^*,2}) \leq N(\varepsilon\overline{N_2^*}^{-1/2},\mathcal{F}_\delta,||.||_{\mathbb{Q}_n^*,2}).$$

The Cauchy-Schwarz inequality also implies

$$\sigma_{1,2}^{*2} \leq \widetilde{\sigma}_n^{*2} = \sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\delta}\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\left(\widetilde{f}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}} = \sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\delta}|\mathbb{P}_n^*\widetilde{f}^2|.$$

Monotonicity of the integral, Points 3 and 4 of Lemma S4.11 and the inequality $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ entail

$$\sqrt{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\left|\frac{1}{n}\sum_{i_1=1}^{n}\varepsilon_{\{i_1\}}\frac{(n-k)!}{(n-1)!}\sum_{(i_2,\ldots,i_k)\in\mathbb{I}_{n,k-1}}\widetilde{f}(\widetilde{Y}_{\boldsymbol{i}^*})\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right|\middle|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k},(\boldsymbol{i}^*)_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\right]$$
$$\leq\mathbb{E}\left[\widetilde{\sigma}_n^* + \int_0^{\widetilde{\sigma}_n^*}\sqrt{\log N\left(\varepsilon/(4\sqrt{N_2^*}),\mathcal{F},||.||_{\mathbb{Q}_n^*,2}\right)}d\varepsilon\middle|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k},(\boldsymbol{i}^*)_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\right]\mathbb{1}\{\overline{N_2^*}>0\}.$$

Furthermore when $\overline{N_2^*}>0$ the following holds:

$$\begin{aligned}
&\int_0^{\widetilde{\sigma}_n^*}\sqrt{\log N\left(\varepsilon/(4\overline{N_2^*}^{1/2}),\mathcal{F},||.||_{\mathbb{Q}_n^*,2}\right)}d\varepsilon\\
=&\int_0^{\widetilde{\sigma}_n^*}\sqrt{\log N\left(\varepsilon||F||_{\mathbb{Q}_n^*,2}/(4\overline{N_2^*}^{1/2}||F||_{\mathbb{Q}_n^*,2}),\mathcal{F},||.||_{\mathbb{Q}_n^*,2}\right)}d\varepsilon\\
=&4\overline{N_2^*}^{1/2}||F||_{\mathbb{Q}_n^*,2}\int_0^{\widetilde{\sigma}_n^*/(4\overline{N_2^*}^{1/2}||F||_{\mathbb{Q}_n^*,2})}\sqrt{\log N\left(\varepsilon||F||_{\mathbb{Q}_n^*,2},\mathcal{F},||.||_{\mathbb{Q}_n^*,2}\right)}d\varepsilon\\
\leq&4\sqrt{\overline{N_2^*}||F||_{\mathbb{Q}_n^*,2}^2}J_{\mathcal{F}}\left(\frac{\sqrt{\widetilde{\sigma}_n^{*2}}}{4\sqrt{\overline{N_2^*}||F||_{\mathbb{Q}_n^*,2}^2}}\right).
\end{aligned}\tag{4.32}$$

Let $A_n = \mathbb{P}\left(\overline{N_2^*}>0\mid(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)$. Relation (4.32), Lemma S4.10, Jensen's inequality and the fact that $\mathbb{E}\left(\widetilde{\sigma}_n^{*2}|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k},\overline{N_2^*}>0\right)A_n = \mathbb{E}\left(\widetilde{\sigma}_n^{*2}|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)$ and $\mathbb{E}\left(N^*||F||_{\mathbb{Q}_n^*,2}^2|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k},\overline{N_2^*}>0\right)A_n = \frac{1}{n^k}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}N_{\boldsymbol{i}}\sum_{\ell=1}^{N_{\boldsymbol{i}}}F^2(Y_{\boldsymbol{i},\ell})$ thus ensure for some constant K_k' that depends on k only

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\left|\widetilde{\mathbb{G}}_n^*f\right|\middle|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right]\\
\leq&K_k'\mathbb{E}\left(\widetilde{\sigma}_n^{*2}|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)^{1/2}\\
&+K_k'\left(\frac{1}{n^k}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}N_{\boldsymbol{i}}\sum_{\ell=1}^{N_{\boldsymbol{i}}}F^2(Y_{\boldsymbol{i},\ell})\right)^{1/2}J_{\mathcal{F}}\left(\frac{\mathbb{E}\left(\widetilde{\sigma}_n^{*2}|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)^{1/2}}{4\left(\frac{1}{n^k}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}N_{\boldsymbol{i}}\sum_{\ell=1}^{N_{\boldsymbol{i}}}F^2(Y_{\boldsymbol{i},\ell})\right)^{1/2}}\right)\sqrt{A_n}.
\end{aligned}$$

Since $\frac{1}{n^k}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} N_{\boldsymbol{i}}\sum_{\ell=1}^{N_{\boldsymbol{i}}} F^2(Y_{\boldsymbol{i},\ell}) \xrightarrow{a.s.} \mathbb{E}\left(N_{\mathbf{1}}\sum_{\ell=1}^{N_{\mathbf{1}}} F^2(Y_{\mathbf{1},\ell})\right)$ and $A_n \leq 1$, we only have to show that

$$\limsup_{n\to\infty} \mathbb{E}\left(\widetilde{\sigma}_n^{*2}|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right) \xrightarrow{a.s.} 0 \text{ as } \delta \downarrow 0.$$

We have:

$$\begin{aligned}
\widetilde{\sigma}_n^{*2} =& \sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\delta} |\mathbb{P}_n^* \widetilde{f}^2| \\
\leq& \sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\delta} \left|\mathbb{P}_n^* \widetilde{f}^2 - \frac{n!}{n^k(n-k)!}\mathbb{P}_n \widetilde{f}^2\right| + \frac{n!}{n^k(n-k)!}\left(\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\delta}\left|\mathbb{P}_n\widetilde{f}^2 - P\widetilde{f}^2\right| + \delta^2\right) \\
\leq& \sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\infty} \left|\mathbb{P}_n^* \widetilde{f}^2 - \frac{n!}{n^k(n-k)!}\mathbb{P}_n \widetilde{f}^2\right| + \sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\infty}\left|\mathbb{P}_n\widetilde{f}^2 - P\widetilde{f}^2\right| + \delta^2
\end{aligned}$$

In the proof of Theorem 4.5.2, we have shown that $\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\infty} |\mu_n^r \widetilde{f}^2 - P\widetilde{f}^2|$ converges in L^1 to 0 for every $r = 1, ..., k$. A similar proof can be used to claim that $\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\infty} |\mathbb{P}_n \widetilde{f}^2 - P\widetilde{f}^2|$ converges in L^1 to 0. A backwards submartingale argument used in the proof of Theorem 4.1.1 ensures that this convergence is almost sure. Because $\frac{n!}{n^k(n-k)!}$ tends to 1, it is sufficient to show that

$$\mathbb{E}\left(\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\infty}\left|\mathbb{P}_n^*\widetilde{f}^2 - \frac{n!}{n^k(n-k)!}\mathbb{P}_n\widetilde{f}^2\right| \mid (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right) \xrightarrow{a.s.} 0.$$

Note that $\mathbb{E}\left(\mathbb{P}_n^*\widetilde{f}^2|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right) = \frac{n!}{n^k(n-k)!}\mathbb{P}_n\widetilde{f}^2$. The symmetrization step in Lemma S4.6 ensures

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\infty}\left|\mathbb{P}_n^*\widetilde{f}^2 - \frac{n!}{n^k(n-k)!}\mathbb{P}_n\widetilde{f}^2\right| \mid (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right] \\
\leq& 4\frac{1}{n^k}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\left(\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}})\right)^2 \mathbb{1}_{\left\{\left(\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}})\right)^2 > M\right\}} \\
&+ kC_k\mathbb{E}\left[\sup_{f\in\mathcal{F}_\infty}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{i_1\}}\left(\widetilde{f}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2 \mathbb{1}_{\left\{\left(\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2 \leq M\right\}}\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right| \middle| (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right],
\end{aligned}$$

for some positive constant C_k that depends on k only.

If $\overline{N_2^*} = 0$,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_\infty}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{i_1\}}\left(\widetilde{f}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2 \mathbb{1}_{\left\{\left(\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2 \leq M\right\}}\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right| \middle| (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}, (\boldsymbol{i}^*)_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\right]$$

is null. Otherwise $\overline{N_2^*} > 0$ and conditional on $\left((\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}, (\boldsymbol{i}^*)_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\right)$, we can consider for every $\eta_1 > 0$ a minimal η_1-covering of $\widetilde{\mathcal{F}}_\infty^2 = \{g = (\widetilde{f}_1 - \widetilde{f}_2)^2 : (\widetilde{f}_1, \widetilde{f}_2) \in \mathcal{F}\times\mathcal{F}\}$ for the seminorm

$$||g||_{M,1}^* = \frac{(n-k)!}{n!}\sum_{i_1=1}^n \left|\sum_{(i_2,...,i_k):\boldsymbol{i}\in\mathbb{I}_{n,k}} g\left(\widetilde{Y}_{\boldsymbol{i}^*}\right)\mathbb{1}_{\left\{\left(\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2\leq M\right\}}\mathbb{1}_{\{i^*\in\mathbb{I}_{n,k}\}}\right|$$

with balls centered in $\mathcal{F}$. This implies

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}_\infty}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{i_1\}}\left(\widetilde{f}\left(\widetilde{Y}_{\boldsymbol{i}^*}\right)\right)^2 \mathbb{1}_{\left\{\left(\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2 \leq M\right\}}\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right| \middle| (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}, (\boldsymbol{i}^*)_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\right] \\
\leq& 4\sqrt{2\log 2N\left(\eta_1, \widetilde{\mathcal{F}}_\infty^2, ||.||_{M,1}^*\right)}M\frac{1}{\sqrt{n}} + \eta_1.
\end{aligned}$$

Remark that for $\widetilde{f} \in \widetilde{\mathcal{F}}_\infty$ with corresponding $f \in \mathcal{F}_\infty$, $||\widetilde{f}^2||^*_{M,1} \leq \overline{N_2^*}||f^2||_{\mathbb{Q}_n^*,1}$ where $||g||_{\mathbb{Q}_n^*,1} = \frac{1}{\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} N_{\boldsymbol{i}^*}^2 \mathbb{1}_{\left\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\right\}}} \sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}} N_{\boldsymbol{i}^*} \sum_{\ell=1}^{N_{\boldsymbol{i}^*}} |g(Y_{\boldsymbol{i}^*,\ell})| \mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}$. Then, for every $\eta > 0$, using Points 1, 2 and 4 of Lemma S4.11 and letting $\eta_1 = 8\eta\overline{N_2^*}||F^2||_{\mathbb{Q}_n^*,1}$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathcal{F}_\infty}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{i_1\}}\left(\widetilde{f}\left(\widetilde{Y}_{\boldsymbol{i}^*}\right)\right)^2\mathbb{1}_{\left\{\left(\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2\leq M\right\}}\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right|\middle|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k},(\boldsymbol{i}^*)_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\right]\\
\leq&4\sqrt{2\log 2\sup_Q N^2\left(\eta||F||_{Q,2},\mathcal{F},||.||_{Q,2}\right)}M\frac{1}{\sqrt{n}}+8\eta\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}N_{\boldsymbol{i}^*}\sum_{\ell=1}^{N_{\boldsymbol{i}^*}}F^2(Y_{\boldsymbol{i}^*,\ell})\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}.
\end{aligned}$$

Integration with respect to $(\boldsymbol{i}^*)_{\boldsymbol{i}\in\mathbb{I}_{n,k}}|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$ leads to

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}_\infty}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{i_1\}}\left(\widetilde{f}\left(\widetilde{Y}_{\boldsymbol{i}^*}\right)\right)^2\mathbb{1}_{\left\{\left(\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2\leq M\right\}}\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right|\middle|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right]\\
\leq&4\sqrt{2\log 2\sup_Q N^2\left(\eta||F||_{Q,2},\mathcal{F},||.||_{Q,2}\right)}M\frac{1}{\sqrt{n}}+8\eta\frac{1}{n^k}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}N_{\boldsymbol{i}}\sum_{\ell=1}^{N_{\boldsymbol{i}}}F^2(Y_{\boldsymbol{i},\ell}).
\end{aligned}$$

We observe $\frac{1}{n^k}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\left(\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}})\right)^2\mathbb{1}_{\left\{\left(\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}})\right)^2>M\right\}}$ and $\frac{1}{n^k}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}N_{\boldsymbol{i}}\sum_{\ell=1}^{N_{\boldsymbol{i}}}F^2(Y_{\boldsymbol{i},\ell})$ converge a.s. to $\mathbb{E}\left(\left(\widetilde{F}(\widetilde{Y}_{\mathbf{1}})\right)^2\mathbb{1}_{\left\{\left(\widetilde{F}(\widetilde{Y}_{\mathbf{1}})\right)^2>M\right\}}\right)$ and $\mathbb{E}\left(N_{\mathbf{1}}\sum_{\ell=1}^{N_{\mathbf{1}}}F^2(Y_{\mathbf{1},\ell})\right)$ by almost sure convergence of the sample mean of jointly exchangeable arrays [66] and ergodicity of dissociated arrays [96] or Theorem 4.1 for a class $\mathcal{F}$ reduced to a singleton. Choosing M and η arbitrarily small, we deduce that for $n\to\infty$

$$\mathbb{E}\left(\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\infty}\left|\mathbb{P}_n^*\widetilde{f}^2-\frac{n!}{n^k(n-k)!}\mathbb{P}_n\widetilde{f}^2\right|\mid(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)\xrightarrow{a.s.}0.$$

4.7.2.3 Theorem 4.6

We recall that under multiway clustering, $n_1,...,n_k$ are all indexed by an index m, though we most often leave this dependence implicit hereafter. They also satisfy, as $m\to\infty$, $\underline{n}=\min(n_1,...,n_k)\to\infty$ and $\underline{n}/n_k\to\lambda_j$.

4.7.2.3.1 Uniform law of large numbers

We show hereafter that

$$\sup_{\mathcal{F}}\left|\widetilde{\mathbb{P}}_{\boldsymbol{n}}f-\widetilde{P}f\right|\xrightarrow{L^1,\text{a.s.}}0. \tag{4.33}$$

First, we have $\sup_{f\in\mathcal{F}}|\widetilde{\mathbb{P}}_{\boldsymbol{n}}f-\widetilde{P}f|=\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}}|\mathbb{P}_{\boldsymbol{n}}\widetilde{f}-P\widetilde{f}|$. Next, the triangle inequality and the symmetrization Lemma S4.5 for the class $\mathcal{G}=\left\{\widetilde{f}\mathbb{1}_{\{\widetilde{F}\leq M\}}:\widetilde{f}\in\widetilde{\mathcal{F}}\right\}$ and Φ=Id ensure that for every $M>0$

$$\begin{aligned}
\mathbb{E}\left[\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}}\left|\mathbb{P}_{\boldsymbol{n}}\widetilde{f}-P\widetilde{f}\right|\right]\leq&2\mathbb{E}\left[\widetilde{F}(\widetilde{Y}_{\mathbf{1}})\mathbb{1}_{\{\widetilde{F}(\widetilde{Y}_{\mathbf{1}})>M\}}\right]\\
&+2\sum_{\boldsymbol{e}\in\cup_{r=1}^k\mathcal{E}_r}\mathbb{E}\left[\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\varepsilon_{\boldsymbol{i}\odot\boldsymbol{e}}\widetilde{f}\left(\widetilde{Y}_{\boldsymbol{i}}\right)\mathbb{1}_{\{\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}})\leq M\}}\right|\right].
\end{aligned}$$

For every $\boldsymbol{e}\in\cup_{j=1}^k\mathcal{E}_j$, let

$$||\widetilde{f}||_{\boldsymbol{e},M,1}=\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\boldsymbol{e}\leq\boldsymbol{d}\leq\boldsymbol{n}\odot\boldsymbol{e}}\left|\sum_{\mathbf{1}-\boldsymbol{e}\leq\boldsymbol{d}'\leq\boldsymbol{n}\odot(\mathbf{1}-\boldsymbol{e})}\widetilde{f}\left(\widetilde{Y}_{\boldsymbol{i}}\right)\mathbb{1}_{\left\{\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}})\leq M\right\}}\right|.$$

Using the same steps as in Part 1 of the proof of Theorem 4.5, we get for every $\boldsymbol{e} \in \cup_{j=1}^{k} \mathcal{E}_j$, every $M > 0$ and every possibly random $\eta_1 \geq 0$,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\varepsilon_{\boldsymbol{i}\odot\boldsymbol{e}}\widetilde{f}\left(\widetilde{Y}_{\boldsymbol{i}}\right)\mathbb{1}_{\{\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}})\leq M\}}\right|\right] \\
\leq&\mathbb{E}\left[\sqrt{2\log 2N\left(\eta_1,\mathcal{F},||.||_{\boldsymbol{e},M,1}\right)}M\frac{1}{\sqrt{\prod_{j=1}^{k}n_j\mathbb{1}\{e_j=1\}}}+\eta_1\middle|\overline{N_1}>0\right]\mathbb{P}\left(\overline{N_1}>0\right),
\end{aligned}$$

with $\overline{N_1} = \frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} N_{\boldsymbol{i}}$. Observe that $||\widetilde{f}||_{\boldsymbol{e},M,1} \leq \overline{N_1}||f||_{\mathbb{Q}_{\boldsymbol{n}},1} = \frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\sum_{\ell=1}^{N_{\boldsymbol{i}}}|f(Y_{\boldsymbol{i},\ell})|$ where

$$||f||_{\mathbb{Q}_{\boldsymbol{n}},1} = \frac{1}{\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}N_{\boldsymbol{i}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\sum_{\ell=1}^{N_{\boldsymbol{i}}}|f(Y_{\boldsymbol{i},\ell})|.$$

Letting $\eta_1 = \eta\overline{N_1}||F||_{\mathbb{Q}_{\boldsymbol{n}},1}$, we can follow the proof of Point 1 in Theorem 4.5 to conclude that $\mathbb{E}\left[\sup_{\widetilde{\mathcal{F}}}|\mathbb{P}_{\boldsymbol{n}}\widetilde{f} - P\widetilde{f}|\right]$ tends to 0 as $m \to \infty$.

We now turn to proving almost sure convergence. Let $\Sigma_{\boldsymbol{n}}$ be the σ-algebra generated by $\mathcal{H}_{\boldsymbol{n}}$ the set of functions g from $\mathcal{D}^{\mathbb{N}^{+k}}$ to $\mathbb{R}$ that are invariant by the action of any $(\pi_1, ..., \pi_k)$, with π_r any permutation on $\mathbb{N}^+$ such that $\pi_r(j) = j$ if $j \geq n_r$ for $r = 1, ..., k$:

$$g\left(\left(\widetilde{Y}_{\boldsymbol{i}}\right)_{\boldsymbol{i}\in\mathbb{N}^{+k}}\right) = g\left(\left(\widetilde{Y}_{\pi_1(i_1),...,\pi_k(i_k)}\right)_{\boldsymbol{i}\in\mathbb{N}^{+k}}\right).$$

For every $\boldsymbol{n}' \geq \boldsymbol{n}$, $\boldsymbol{n}' \neq \boldsymbol{n}$, let $\mathbb{J}_{\boldsymbol{n},\boldsymbol{n}'} = \mathbb{I}_{n_1',n_1'-n_1} \times ... \times \mathbb{I}_{n_k',n_k'-n_k}$. Then, for every $q = (q_1, ..., q_k) \in \mathbb{J}_{\boldsymbol{n},\boldsymbol{n}'}$, let

$$\mathbb{P}^q_{\boldsymbol{n},\boldsymbol{n}'}\widetilde{f} = \frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}'}\widetilde{f}(\widetilde{Y}_{\boldsymbol{i}})\mathbb{1}_{\{i_1\notin\{q_1\},...,i_k\notin\{q_k\}\}}.$$

We observe that for every $\boldsymbol{n}, \boldsymbol{n}', q$,

$$\mathbb{E}\left(\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}}\left|\mathbb{P}^q_{\boldsymbol{n},\boldsymbol{n}'}\widetilde{f} - P\widetilde{f}\right||\Sigma_{\boldsymbol{n}'}\right) = \mathbb{E}\left(\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}}\left|\mathbb{P}_{\boldsymbol{n}}\widetilde{f} - P\widetilde{f}\right||\Sigma_{\boldsymbol{n}'}\right).$$

Moreover,

$$\begin{aligned}
\sum_{q\in\mathbb{J}_{\boldsymbol{n},\boldsymbol{n}'}}\mathbb{P}^q_{\boldsymbol{n},\boldsymbol{n}'}\widetilde{f} =& \frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}'}\widetilde{f}(\widetilde{Y}_{\boldsymbol{i}})\sum_{q\in\mathbb{J}_{\boldsymbol{n},\boldsymbol{n}'}}\mathbb{1}_{\{i_1\notin\{q_1\},...,i_k\notin\{q_k\}\}} \\
=& \prod_{j=1}^{k}\frac{(n_j'-1)!}{n_j!}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}'}\widetilde{f}(\widetilde{Y}_{\boldsymbol{i}}).
\end{aligned}$$

and next, $\mathbb{P}_{\boldsymbol{n}'}\widetilde{f} = \left(\prod_{j=1}^{k}\frac{n_j!}{n_j'!}\right)\sum_{q\in\mathbb{J}_{\boldsymbol{n},\boldsymbol{n}'}}\mathbb{P}^q_{\boldsymbol{n},\boldsymbol{n}'}\widetilde{f} = \frac{1}{|\mathbb{J}_{\boldsymbol{n},\boldsymbol{n}'}|}\sum_{q\in\mathbb{J}_{\boldsymbol{n},\boldsymbol{n}'}}\mathbb{P}^q_{\boldsymbol{n},\boldsymbol{n}'}\widetilde{f}$. Furthermore,

$$\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}}\left|\mathbb{P}_{\boldsymbol{n}'}\widetilde{f} - P\widetilde{f}\right| = \mathbb{E}\left(\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}}\left|\mathbb{P}_{\boldsymbol{n}'}\widetilde{f} - P\widetilde{f}\right||\Sigma_{\boldsymbol{n}'}\right).$$

This last equality, combined with those just above and the triangle inequality give

$$\begin{aligned}
\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}}\left|\mathbb{P}_{\boldsymbol{n}'}\widetilde{f} - P\widetilde{f}\right| &\leq \frac{1}{|\mathbb{J}_{\boldsymbol{n},\boldsymbol{n}'}|}\sum_{q\in\mathbb{J}_{\boldsymbol{n},\boldsymbol{n}'}}\mathbb{E}\left(\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}}\left|\mathbb{P}^q_{\boldsymbol{n},\boldsymbol{n}'}\widetilde{f} - P\widetilde{f}\right||\Sigma_{\boldsymbol{n}'}\right) \\
&= \mathbb{E}\left(\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}}\left|\mathbb{P}_{\boldsymbol{n}}\widetilde{f} - P\widetilde{f}\right||\Sigma_{\boldsymbol{n}'}\right).
\end{aligned}$$

Then considering $\boldsymbol{n} = (n_1(m), ..., n_k(m))$ and $\boldsymbol{n}' = (n_1(m+1), ..., n_k(m+1))$, we deduce from the almost sure convergence of backwards submartingales that $\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}}\left|\mathbb{P}_{\boldsymbol{n}'}\widetilde{f} - P\widetilde{f}\right|$ converges almost surely to 0 when m tends to infinity.

4.7.2.3.2 Uniform central limit theorem

First step: pointwise weak convergence

To prove the pointwise weak convergence, the line of reasoning is the same as what we resorted to in the first step of the proof of Theorem 4.1.2: for every $f \in \mathcal{F}$, we need to find a suitable L_2-approximation of $\mathbb{G}_{\boldsymbol{n}} f$, denoted $H_1 f$, *i.e* as $m \to +\infty$ $H_1 f$ must satisfy $\mathbb{E}\left[\left|\mathbb{G}_{\boldsymbol{n}} f - H_1 f\right|^2\right] = o(1)$ and $H_1 f \xrightarrow[n\to+\infty]{d} \mathcal{N}(0, K(f,f))$. We pick $H_1 f = \sum_{\boldsymbol{e}\in\mathcal{E}_1} \sum_{\mathbf{1}\leq \boldsymbol{i}\leq \boldsymbol{n}} \mathbb{E}\left[\mathbb{G}_{\boldsymbol{n}} f \mid |U_{\boldsymbol{i}\odot\boldsymbol{e}}\right]$, where $(U_{\boldsymbol{i}\odot\boldsymbol{e}})_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n},\boldsymbol{e}\in\mathcal{E}_1}$ are i.i.d terms that appear in the AHK representation of $(\widetilde{Y}_{\boldsymbol{i}})_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}$. Let $\boldsymbol{i}^r$ be a vector with all its entries equal to one except the r-th one, which is equal to i_r. The AHK representation ensures

$$\begin{aligned} H_1(f) &= \sum_{\boldsymbol{e}\in\mathcal{E}_1} \sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} \mathbb{E}\left[\mathbb{G}_{\boldsymbol{n}} f \mid U_{\boldsymbol{i}\odot\boldsymbol{e}}\right] \\ &= \sum_{r=1}^{k} \frac{\sqrt{n}}{n_r} \sum_{i_r=1}^{n_r} \left(\mathbb{E}\left[\widetilde{f}\left(\widetilde{Y}_{\boldsymbol{i}^r}\right) \mid U_{i_r}\right] - \mathbb{E}\left[\widetilde{f}(\widetilde{Y}_{\mathbf{1}})\right]\right) \\ &\xrightarrow[n\to+\infty]{d} \mathcal{N}(0, K(f,f)). \end{aligned}$$

The convergence in distribution comes from the standard central limit theorem applied for each $\boldsymbol{e} \in \mathcal{E}_1$ separately, the mutual independence of terms across $\boldsymbol{e} \in \mathcal{E}_1$ in the previous expression and the fact that $\sqrt{\underline{n}/n_r} \to \sqrt{\lambda_r}$.

To conclude that $\mathbb{G}_{\boldsymbol{n}} f \xrightarrow[n\to+\infty]{d} \mathcal{N}(0, K(f,f))$ as $m \to +\infty$, we rely on the weak convergence of $H_1 f$ and Section C.2.1 in [50]. The main step there amounts to showing that $\lim_{m\to+\infty} \mathbb{V}(H_1 f)/\mathbb{V}(\mathbb{G}_{\boldsymbol{n}} f) = 1$.

Second step: asymptotic equicontinuity

Following the same reasoning as in the proof of Part 2 of Theorem 4.5, with the symmetrization lemma S4.5 instead of Lemma 4.2, we have

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta} \left|\widetilde{\mathbb{G}}_{\boldsymbol{n}} f\right|\right] = \mathbb{E}\left[\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\delta} \left|\mathbb{G}_{\boldsymbol{n}} \widetilde{f}\right|\right] = O\left(\mathbb{E}\left(\int_0^{\widetilde{\sigma}_{\boldsymbol{n}}} \sqrt{\log 2N\left(\varepsilon, \widetilde{\mathcal{F}}_\delta, ||.||_{\mu_{\boldsymbol{n}},2}\right)} d\varepsilon\right)\right),$$

where $\mu_{\boldsymbol{n}} = \frac{1}{\Pi_{\boldsymbol{n}}} \sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} \delta_{(N_{\boldsymbol{i}}, (Y_{\boldsymbol{i},\ell})_{N_{\boldsymbol{i}}\geq\ell\geq 1})}$. $||\widetilde{f}||^2_{\mu_{\boldsymbol{n}},2}$ and $\widetilde{\sigma}^2_{\boldsymbol{n}}$ are defined in the same way as in the proof of Part 2 of Theorem 4.5 (with $\mu_{\boldsymbol{n}}$ instead of μ_n). Still following this proof, we obtain

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta} \left|\widetilde{\mathbb{G}}_{\boldsymbol{n}} f\right|\right] \leq O\left[\mathbb{E}\left(\widetilde{\sigma}^2_{\boldsymbol{n}}\right)^{1/2} + \mathbb{E}\left(N_{\mathbf{1}} \sum_{\ell=1}^{N_{\mathbf{1}}} F^2(Y_{\mathbf{1},\ell})\right)^{1/2} J_{\mathcal{F}}\left(\frac{\mathbb{E}\left(\widetilde{\sigma}^2_{\boldsymbol{n}}\right)^{1/2}}{4\mathbb{E}\left(N_{\mathbf{1}} \sum_{\ell=1}^{N_{\mathbf{1}}} F^2(Y_{\mathbf{1},\ell})\right)^{1/2}}\right)\right].$$

Recalling that $\mathbb{E}\left(\widetilde{\sigma}^2_{\boldsymbol{n}}\right) \leq \mathbb{E}\left[\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\infty} \left|\mathbb{P}_{\boldsymbol{n}} \widetilde{f}^2 - \mathbb{P}\widetilde{f}^2\right|\right] + \delta^2$, we can follow the end of the asymptotic equicontinuity proof of Part 2 of Theorem 4.5 with obvious minor changes to conclude.

Third step: total boundedness

The proof of the total boundedness follows the same lines as in the proof of Theorem 4.5.2 with μ_n, $\overline{N_2}$ and $\mathbb{Q}_n$ replaced respectively by $\mu_{\boldsymbol{n}}$, $\frac{1}{\Pi_{\boldsymbol{n}}} \sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} N_{\boldsymbol{i}}^2$ and $\mathbb{Q}_{\boldsymbol{n}} = \frac{1}{\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} N_{\boldsymbol{i}}^2} \sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} N_{\boldsymbol{i}} \sum_{\ell=1}^{N_{\boldsymbol{i}}} \delta_{Y_{i,\ell}}$.

4.7.2.3.3 Convergence of the bootstrap process

As previously, we only have to prove the pointwise convergence and the asymptotic equicontinuity.

First step: pointwise convergence

Let $\boldsymbol{i}^* = (i_1^*, ..., i_k^*)$ denote the cell obtained by sampling i_j^* with replacement in $1, ..., n_j$ for every $j = 1, ..., k$.

We have the almost-sure representation

$$\boldsymbol{i}^* = (F^{-1}_{n_1}[U^*_{(i_1,0,...,0)}], ..., F^{-1}_{n_k}[U^*_{(0,...,0,i_k)}]),$$

with $(U_A^*)_{A\in\mathbb{N}^k}$ a family of i.i.d. uniform random variables and $F_{n_j}^{-1}$ the quantile function of the discrete uniform distribution on $\{1,...,n_j\}$. Conditional on the data $(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{N}^{+k}}$, we can thus follow an approach similar to the one we used in the jointly exchangeable case. Let $H_1^* f = \sum_{\boldsymbol{e}\in\mathcal{E}_1}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} \mathbb{E}\left[\widetilde{\mathbb{G}}_{\boldsymbol{n}}^* f|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{N}^{+k}}, U^*_{\boldsymbol{i}\odot\boldsymbol{e}}\right]$ and $h(\boldsymbol{i}) = \widetilde{f}(\widetilde{Y}_{\boldsymbol{i}})$. $H_1^* f$ can also be written

$$\sqrt{\underline{n}}\sum_{r=1}^{k}\left(\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} h(i_1,...,i_{r-1},i_r^*,i_{r+1},...,i_k) - \widetilde{\mathbb{P}}_{\boldsymbol{n}} f\right).$$

We first show that $\mathbb{E}\left[\left(\widetilde{\mathbb{G}}_{\boldsymbol{n}}^* f - H_1^* f\right)^2 \mid (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{N}^{+k}}\right] = o_{a.s}(1)$. Expanding the square in the previous formula gives

$$\begin{aligned}
&\mathbb{E}\left[\left(\widetilde{\mathbb{G}}_{\boldsymbol{n}}^* f - H_1^* f\right)^2 |(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{N}^{+k}}\right]\\
=&\underline{n}\Bigg\{\mathbb{E}\left[\left(\sum_{r=1}^{k}\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} h(i_1,...,i_{r-1},i_r^*,i_{r+1},...,i_k)\right)^2 \mid (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{N}^{+k}}\right]\\
&-2\mathbb{E}\left[\left(\sum_{r=1}^{k}\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} h(i_1,...,i_{r-1},i_r^*,i_{r+1},...,i_k)\right)\widetilde{\mathbb{P}}_{\boldsymbol{n}}^* f \mid (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{N}^{+k}}\right]\\
&+\mathbb{E}\left[\left(\widetilde{\mathbb{P}}_{\boldsymbol{n}}^* f\right)^2 \mid (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{N}^{+k}}\right] - (k-1)^2\left(\widetilde{\mathbb{P}}_{\boldsymbol{n}} f\right)^2\Bigg\}.
\end{aligned}$$

Let $A_{\boldsymbol{n}} = \sum_{r=1}^{k}\frac{1}{\Pi_{\boldsymbol{n}}^2}\sum_{\substack{\mathbf{1}\leq\boldsymbol{i},\boldsymbol{i}'\leq\boldsymbol{n}\\ i_r=i_r'}} h(\boldsymbol{i})h(\boldsymbol{i}')$. We can show

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{r=1}^{k}\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} h(i_1,...,i_{r-1},i_r^*,i_{r+1},...,i_k)\right)^2 \mid (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{N}^{+k}}\right]\\
&=\left(\widetilde{\mathbb{P}}_{\boldsymbol{n}} f\right)^2\left(\sum_{r=1}^{k}\frac{(n_r-1)}{n_r}+k(k-1)\right)+A_{\boldsymbol{n}},\\
&\mathbb{E}\left[\left(\sum_{r=1}^{k}\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} h(i_1,...,i_{r-1},i_r^*,i_{r+1},...,i_k)\right)\widetilde{\mathbb{P}}_{\boldsymbol{n}}^* f \mid (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{N}^{+k}}\right]\\
&=\left(\widetilde{\mathbb{P}}_{\boldsymbol{n}} f\right)^2\sum_{r=1}^{k}\frac{(n_r-1)}{n_r}+A_{\boldsymbol{n}},
\end{aligned}$$

and $\mathbb{E}\left[\left(\widetilde{\mathbb{P}}_{\boldsymbol{n}}^* f\right)^2 \mid (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{N}^{+k}}\right] = \frac{\Pi_{j=1}^{k}(n_j-1)}{\Pi_{\boldsymbol{n}}}\left(\widetilde{\mathbb{P}}_{\boldsymbol{n}} f\right)^2 + B_{\boldsymbol{n}}$, where

$$B_{\boldsymbol{n}} = \frac{1}{\Pi_{\boldsymbol{n}}}\sum_{r=1}^{k}\sum_{\boldsymbol{e}\in\mathcal{E}_r}\frac{\prod_{1\leq j\leq k:e_j=0}(n_j-1)}{\prod_{1\leq j\leq k:e_j=1} n_j\left(\prod_{1\leq j\leq k:e_j=0} n_j\right)^2}\sum_{\substack{\mathbf{1}\leq\boldsymbol{i},\boldsymbol{i}'\leq\boldsymbol{n}\\ i_j=i_j'\forall j:e_j=1}} h(\boldsymbol{i})h(\boldsymbol{i}').$$

For every $\boldsymbol{e}\in\cup_{r=2}^{k}\mathcal{E}_r$, we can write the following decomposition

$$\sum_{\substack{\mathbf{1}\leq\boldsymbol{i},\boldsymbol{i}'\leq\boldsymbol{n}\\ i_j=i_j'\forall j:e_j=1}} h(\boldsymbol{i})h(\boldsymbol{i}') = \sum_{\substack{\boldsymbol{e}'\in\cup_{r=1}^{k}\mathcal{E}_r\\ e_j'=1 \text{ if } e_j=1}}\sum_{(\boldsymbol{i},\boldsymbol{i}')\in\mathcal{I}_{\boldsymbol{n},\boldsymbol{e}'}} h(\boldsymbol{i})h(\boldsymbol{i}'),$$

with $\mathcal{I}_{\boldsymbol{n},\boldsymbol{e}'} = \{(\boldsymbol{i},\boldsymbol{i}') : \mathbf{1}\leq\boldsymbol{i},\boldsymbol{i}'\leq\boldsymbol{n},\ i_r=i_r' \text{ if } e_r'=1 \text{ and } i_r\neq i_r' \text{ otherwise}\}$. Applying Lemma S4.9, we

conclude that for every $\boldsymbol{e} \in \cup_{r=2}^{k} \mathcal{E}_r$,

$$\sum_{\substack{\mathbf{1} \leq \boldsymbol{i}, \boldsymbol{i}' \leq \boldsymbol{n} \\ i_j = i'_j \forall j : e_j = 1}} h(\boldsymbol{i}) h(\boldsymbol{i}') = O_{\text{a.s}} \left(\Pi_{\boldsymbol{n}} \prod_{1 \leq j \leq k : e_j = 0} (n_j - 1) \right),$$

$$B_{\boldsymbol{n}} = \sum_{r=1}^{k} \frac{n_r \prod_{1 \leq j \leq k : j \neq r} (n_j - 1)}{\Pi_{\boldsymbol{n}}} \frac{1}{\Pi_{\boldsymbol{n}}^2} \sum_{\substack{\mathbf{1} \leq \boldsymbol{i}, \boldsymbol{i}' \leq \boldsymbol{n} \\ i_r = i'_r}} h(\boldsymbol{i}) h(\boldsymbol{i}') + O_{\text{a.s}}(\underline{n}^{-2}).$$

By combining all those elements, we obtain

$$\mathbb{E}\left[\left(\widetilde{\mathbb{G}}_{\boldsymbol{n}}^* f - H_1^* f \right)^2 | (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{N}^{+k}} \right] = \underline{n} \Bigg\{ \frac{1}{\Pi_{\boldsymbol{n}}^2} \sum_{r=1}^{k} \left(\frac{n_r \prod_{1 \leq j \leq k : j \neq r} (n_j - 1)}{\Pi_{\boldsymbol{n}}} - 1 \right) \sum_{\substack{\mathbf{1} \leq \boldsymbol{i} \leq \boldsymbol{n} \\ \mathbf{1} \leq \boldsymbol{i}' \leq \boldsymbol{n} \\ i_r = i'_r}} h(\boldsymbol{i}) h(\boldsymbol{i}')$$

$$+ \left(\widetilde{\mathbb{P}}_{\boldsymbol{n}} f \right)^2 \left(\frac{\prod_{1 \leq j \leq k : j \neq r} (n_j - 1)}{\Pi_{\boldsymbol{n}}} - 1 + \sum_{r=1}^{k} \frac{1}{n_r} \right) + O_{\text{a.s}} \left(\underline{n}^{-2} \right) \Bigg\}.$$

Noting that $\frac{n_r \prod_{1 \leq j \leq k : j \neq r} (n_j - 1)}{\Pi_{\boldsymbol{n}}} - 1 = o(1)$, $\frac{\prod_{1 \leq j \leq k : j \neq r} (n_j - 1)}{\Pi_{\boldsymbol{n}}} - 1 + \sum_{r=1}^{k} \frac{1}{n_r} = O\left(\underline{n}^{-2} \right)$ and $\frac{1}{\Pi_{\boldsymbol{n}}^2} \sum_{\substack{\mathbf{1} \leq \boldsymbol{i} \leq \boldsymbol{n} \\ \mathbf{1} \leq \boldsymbol{i}' \leq \boldsymbol{n} \\ i_r = i'_r}} h(\boldsymbol{i}) h(\boldsymbol{i}') = O_{\text{a.s}}(\underline{n}^{-1})$, again by Lemma S4.9, we conclude that

$$\mathbb{E}\left[\left(\widetilde{\mathbb{G}}_{\boldsymbol{n}}^* f - H_1^* f \right)^2 | (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{N}^{+k}} \right] = o_{a.s}(1).$$

To prove the asymptotic normality of $H_1 f$ conditional on $(N_{\boldsymbol{i}}, (Y_{\boldsymbol{i},\ell})_{1 \leq \ell \leq N_{\boldsymbol{i}}})_{\boldsymbol{i} \in \mathbb{N}^{+k}}$, we remark that

$$H_1 f = \sum_{r=1}^{k} \sqrt{\frac{\underline{n}}{n_r}} \sum_{i_r=1}^{n_r} \frac{z_{m,r,i_r}^*}{\sqrt{n_r}},$$

where $z_{m,r,i_r}^* = \frac{1}{\prod_{1 \leq j \leq k : j \neq r} n_j} \sum_{i_j = 1, ..., n_j, \forall j \neq r} \left(h(i_1, ..., i_{r-1}, i_r^*, i_{r+1}, ..., i_k) - \widetilde{\mathbb{P}}_{\boldsymbol{n}} f \right)$. , For every $r = 1, ..., k$, $\left(z_{m,r,i_r}^* \right)_{i_r = 1 ... n_r}$ is an i.i.d. sequence of centered random variables conditional on $(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{N}^{+k}}$ with a distribution that depends on m. Since

$$\mathbb{V}\left(z_{m,r,1}^* \mid (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{N}^{+k}} \right) = \frac{1}{n_r \prod_{1 \leq j \leq k : j \neq r} n_j^2} \sum_{\substack{\mathbf{1} \leq \boldsymbol{i}, \boldsymbol{i}' \leq \boldsymbol{n} \\ i_r = i'_r}} h(\boldsymbol{i}) h(\boldsymbol{i}') - \left(\widetilde{\mathbb{P}}_{\boldsymbol{n}} f \right)^2,$$

we can conclude thanks to Point 1 of Theorem 4.6 and Lemma S4.9 that $\mathbb{V}\left(z_{m,r,1}^* \mid (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{N}^{+k}} \right) \xrightarrow{a.s.} \mathbb{E}\left[h(\mathbf{1}) h(\mathbf{2}_r) \right] - \mathbb{E}\left[h(\mathbf{1}) \right]^2 = \mathbb{C}ov\left(h(\mathbf{1}), h(\mathbf{2}_r) \right) = V_r$. It is not difficult to see that arguments similar to those of substeps 2 and 3 of Section 4.7.1.3 apply. Then, for every $r = 1, ..., k$ and every $t \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(\mathrm{i} t \sum_{i_r=1}^{n_r} \frac{z_{m,r,i_r}^*}{\sqrt{n_r}} \right) \mid (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{N}^{+k}} \right] \xrightarrow{a.s.} \exp\left(-\frac{t^2 V_r}{2} \right).$$

The continuous mapping theorem, the fact that $\frac{\underline{n}}{n_r} \to \lambda_r$ and the mutual independence between the k sequences $\left(z_{m,r,i_r}^* \right)_{i_r = 1 ... n_r}$ $(r = 1, ..., k)$ conditional on the data imply that

$$\mathbb{E}\left[\exp\left(\mathrm{i} t H_1 f \right) \mid (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{N}^{+k}} \right] = \prod_{r=1}^{k} \mathbb{E}\left[\exp\left(\mathrm{i} \sqrt{\frac{\underline{n}}{n_r}} t \sum_{i_r=1}^{n_r} \frac{z_{m,r,i_r}^*}{\sqrt{n_r}} \right) \mid (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{N}^{+k}} \right]$$

$$\xrightarrow{a.s.} \exp\left(-\frac{t^2 \sum_{r=1}^{k} \lambda_r V_r}{2} \right).$$

The result follows.

Second step: asymptotic equicontinuity

First, we have

$$(\boldsymbol{i}^*)_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} = \left(F_{n_1}^{-1}[U^*_{(i_1,0,\ldots,0)}],\ldots,F_{n_k}^{-1}[U^*_{(0,\ldots,0,i_k)}]\right)_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}.$$

This representation ensures that the symmetrization Lemma S4.5 for the class $\widetilde{\mathcal{F}}_\delta$ and Φ=Id is valid. We notice that the representation is "simplified" as only terms associated with $\boldsymbol{e}\in\mathcal{E}_1$ appear. This implies that the telescoping argument in the proof of Lemma S4.5 only has to be undertaken over $\mathcal{E}_1$. The following symmetrization inequality thus holds:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\left|\widetilde{\mathbb{G}}^*_{\boldsymbol{n}}f\right| \mid \left(\widetilde{Y}_{\boldsymbol{i}'}\right)_{\boldsymbol{i}'\geq\mathbf{1}}\right] \leq 2\sum_{\boldsymbol{e}\in\mathcal{E}_1}\mathbb{E}\left[\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\delta}\left|\frac{1}{\sqrt{\Pi_{\boldsymbol{n}}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\varepsilon_{\boldsymbol{i}\odot\boldsymbol{e}}\widetilde{f}\left(\widetilde{Y}_{\boldsymbol{i}^*}\right)\right| \mid \left(\widetilde{Y}_{\boldsymbol{i}'}\right)_{\boldsymbol{i}'\geq\mathbf{1}}\right].$$

Let $\overline{N_2^*} = \frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}N^2_{\boldsymbol{i}^*}$. We can see

$$\begin{aligned}
&2\sum_{\boldsymbol{e}\in\mathcal{E}_1}\mathbb{E}\left[\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\delta}\left|\frac{1}{\sqrt{\Pi_{\boldsymbol{n}}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\varepsilon_{\boldsymbol{i}\odot\boldsymbol{e}}\widetilde{f}\left(\widetilde{Y}_{\boldsymbol{i}^*}\right)\right| \mid \left(\widetilde{Y}_{\boldsymbol{i}'}\right)_{\boldsymbol{i}'\geq\mathbf{1}}\right]\\
=&2\sum_{\boldsymbol{e}\in\mathcal{E}_1}\mathbb{E}\left[\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\delta}\left|\frac{1}{\sqrt{\Pi_{\boldsymbol{n}}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\varepsilon_{\boldsymbol{i}\odot\boldsymbol{e}}\widetilde{f}\left(\widetilde{Y}_{\boldsymbol{i}^*}\right)\right| \mid \left(\widetilde{Y}_{\boldsymbol{i}'}\right)_{\boldsymbol{i}'\geq\mathbf{1}}, \overline{N_2^*}>0\right]\mathbb{P}\left(\overline{N_2^*}>0 \mid \left(\widetilde{Y}_{\boldsymbol{i}'}\right)_{\boldsymbol{i}'\geq\mathbf{1}}\right).
\end{aligned}$$

For every $\boldsymbol{e}\in\mathcal{E}_1$, let $r_{\boldsymbol{e}}$ be the position of the unique non-null element of $\boldsymbol{e}$. This allows us to define

$$||\widetilde{f}||^*_{\boldsymbol{e},2} = \frac{1}{n_{r_{\boldsymbol{e}}}}\sum_{i_{r_{\boldsymbol{e}}}=1}^{n_{r_{\boldsymbol{e}}}}\left[\frac{1}{\prod_{j\neq r_{\boldsymbol{e}}}n_j}\sum_{(i_1,\ldots,i_{r_{\boldsymbol{e}}-1},i_{r_{\boldsymbol{e}}+1},\ldots,i_k:\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n})}\sum_{\ell=1}^{N_{\boldsymbol{i}^*}}f(Y_{\boldsymbol{i}^*,\ell})\right]^2,$$

and $\widetilde{\sigma}^*_{\boldsymbol{n},\boldsymbol{e}} = \sup_{f\in\mathcal{F}_\delta}||\widetilde{f}||^*_{\boldsymbol{e},2}$. Then, by Theorem 2.3.6 in [79], we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\left|\widetilde{\mathbb{G}}^*_{\boldsymbol{n}}f\right| \mid \left(\widetilde{Y}_{\boldsymbol{i}'}\right)_{\boldsymbol{i}'\geq\mathbf{1}}\right]\\
\leq&8\sqrt{2}\sum_{\boldsymbol{e}\in\mathcal{E}_1}\frac{1}{\sqrt{n_{r_{\boldsymbol{e}}}}}\mathbb{E}\left[\sqrt{\log 2}\sigma^*_{\boldsymbol{n},\boldsymbol{e}}+\int_0^{\widetilde{\sigma}^*_{\boldsymbol{n},\boldsymbol{e}}}\sqrt{\log N\left(\varepsilon,\widetilde{\mathcal{F}}_\delta,||\cdot||^*_{\boldsymbol{e},2}\right)}d\varepsilon \mid \left(\widetilde{Y}_{\boldsymbol{i}'}\right)_{\boldsymbol{i}'\geq\mathbf{1}}, \overline{N_2^*}>0\right]\\
&\times\mathbb{P}\left(\overline{N_2^*}>0 \mid \left(\widetilde{Y}_{\boldsymbol{i}'}\right)_{\boldsymbol{i}'\geq\mathbf{1}}\right).
\end{aligned}$$

By a convexity argument, we have, for every $\boldsymbol{e}\in\mathcal{E}_1$ $||\widetilde{f}||^*_{\boldsymbol{e},2}\leq\overline{N_2^*}^{1/2}||f||_{\mathbb{Q}^*_{\boldsymbol{n}},2}$, with $||f||^2_{\mathbb{Q}^*_{\boldsymbol{n}},2} = \overline{N_2^*}^{-1}\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}N_{\boldsymbol{i}^*}\sum_{\ell=1}^{N_{\boldsymbol{i}^*}}f(Y_{\boldsymbol{i}^*,\ell})^2$. We also have $\sigma^{*2}_{\boldsymbol{n},\boldsymbol{e}}\leq\widetilde{\sigma}^{*2}_{\boldsymbol{n}}$, with
$\widetilde{\sigma}^{*2}_{\boldsymbol{n}} = \sup_{f\in\mathcal{F}_\delta}\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\left(\sum_{\ell=1}^{N_{\boldsymbol{i}^*}}f(Y_{\boldsymbol{i}^*,\ell})\right)^2$. As a result (using also Points 1 to 4 of Lemma S4.11),

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\left|\widetilde{\mathbb{G}}^*_{\boldsymbol{n}}f\right| \mid \left(\widetilde{Y}_{\boldsymbol{i}'}\right)_{\boldsymbol{i}'\geq\mathbf{1}}\right]\\
\leq&8\sqrt{2}k\frac{1}{\sqrt{\underline{n}}}\mathbb{E}\left[\sqrt{\log 2}\widetilde{\sigma}^*_{\boldsymbol{n}}+\int_0^{\widetilde{\sigma}^*_{\boldsymbol{n}}}\sqrt{\log N\left(\varepsilon,\mathcal{F}_\delta,\overline{N_2^*}^{1/2}||\cdot||_{\mathbb{Q}^*_{\boldsymbol{n}},2}\right)}d\varepsilon \mid \left(\widetilde{Y}_{\boldsymbol{i}'}\right)_{\boldsymbol{i}'\geq\mathbf{1}}, \overline{N_2^*}>0\right]\\
&\times\mathbb{P}\left(\overline{N_2^*}>0 \mid \left(\widetilde{Y}_{\boldsymbol{i}'}\right)_{\boldsymbol{i}'\geq\mathbf{1}}\right)\\
\leq&8\sqrt{2}k\frac{1}{\sqrt{\underline{n}}}\mathbb{E}\left[\sqrt{\log 2}\widetilde{\sigma}^*_{\boldsymbol{n}}+\int_0^{\widetilde{\sigma}^*_{\boldsymbol{n}}}\sqrt{2\log N\left(\varepsilon/4\overline{N_2^*}^{1/2},\mathcal{F},||\cdot||_{\mathbb{Q}^*_{\boldsymbol{n}},2}\right)}d\varepsilon \mid \left(\widetilde{Y}_{\boldsymbol{i}'}\right)_{\boldsymbol{i}'\geq\mathbf{1}}, \overline{N_2^*}>0\right]\\
&\times\mathbb{P}\left(\overline{N_2^*}>0 \mid \left(\widetilde{Y}_{\boldsymbol{i}'}\right)_{\boldsymbol{i}'\geq\mathbf{1}}\right).
\end{aligned}$$

The same arguments as in (4.32) and in the paragraph that follows this equation lead us to

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\left|\widetilde{\mathbb{G}}^*_{\boldsymbol{n}}f\right| \mid \left(\widetilde{Y}_{\boldsymbol{i}'}\right)_{\boldsymbol{i}'\geq\mathbf{1}}\right] \\
\leq& 8\sqrt{2}k\left\{\sqrt{\log 2}\sqrt{\mathbb{E}\left[\widetilde{\sigma}^{*2}_{\boldsymbol{n}} \mid \left(\widetilde{Y}_{\boldsymbol{i}'}\right)_{\boldsymbol{i}'\geq\mathbf{1}}\right]}\right. \\
&\left. +4\sqrt{\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} N_{\boldsymbol{i}}\sum_{\ell=1}^{N_{\boldsymbol{i}}}F^2(Y_{\boldsymbol{i},\ell})}J_{\mathcal{F}}\left(\frac{\sqrt{\mathbb{E}\left[\widetilde{\sigma}^{*2}_{\boldsymbol{n}} \mid \left(\widetilde{Y}_{\boldsymbol{i}'}\right)_{\boldsymbol{i}'\geq\mathbf{1}}\right]}}{4\sqrt{\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} N_{\boldsymbol{i}}\sum_{\ell=1}^{N_{\boldsymbol{i}}}F^2(Y_{\boldsymbol{i},\ell})}}\right)\right\}.
\end{aligned}$$

Since $\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} N_{\boldsymbol{i}}\sum_{\ell=1}^{N_{\boldsymbol{i}}}F^2(Y_{\boldsymbol{i},\ell}) \xrightarrow{a.s.} \mathbb{E}\left(N_{\mathbf{1}}\sum_{\ell=1}^{N_{\mathbf{1}}}F^2(Y_{\mathbf{1},\ell})\right) > 0$, we only have to show that

$$\limsup_{n\to\infty} E\left[\widetilde{\sigma}^{*2}_{\boldsymbol{n}} \mid \left(\widetilde{Y}_{\boldsymbol{i}'}\right)_{\boldsymbol{i}'\geq\mathbf{1}}\right] \xrightarrow{a.s.} 0 \text{ as } \delta\downarrow 0.$$

We have:

$$\widetilde{\sigma}^{*2}_{\boldsymbol{n}} = \sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\delta}|\mathbb{P}^*_{\boldsymbol{n}}\widetilde{f}^2| \leq \sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\infty}\left|\mathbb{P}^*_{\boldsymbol{n}}\widetilde{f}^2 - \mathbb{P}_{\boldsymbol{n}}\widetilde{f}^2\right| + \sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\infty}\left|\mathbb{P}_{\boldsymbol{n}}\widetilde{f}^2 - P\widetilde{f}^2\right| + \delta^2.$$

In the proof of Point 2 of Theorem 4.6, we have shown $\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\infty}|\mathbb{P}_{\boldsymbol{n}}\widetilde{f}^2 - P\widetilde{f}^2| \xrightarrow{a.s.} 0$. It is therefore sufficient to show

$$\mathbb{E}\left(\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\infty}\left|\mathbb{P}^*_{\boldsymbol{n}}\widetilde{f}^2 - \mathbb{P}_{\boldsymbol{n}}\widetilde{f}^2\right| \mid (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\geq\mathbf{1}}\right) \xrightarrow{a.s.} 0.$$

The symmetrization argument we used to control $\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\left|\widetilde{\mathbb{G}}^*_{\boldsymbol{n}}f\right| \mid \left(\widetilde{Y}_{\boldsymbol{i}'}\right)_{\boldsymbol{i}'\geq\mathbf{1}}\right]$ still applies and gives

$$\begin{aligned}
\mathbb{E}\left[\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\infty}\left|\mathbb{P}^*_{\boldsymbol{n}}\widetilde{f}^2 - \mathbb{P}_{\boldsymbol{n}}\widetilde{f}^2\right| \mid (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\geq\mathbf{1}}\right] \leq& 4\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\in\boldsymbol{n}}\left(\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}})\right)^2\mathbb{1}_{\left\{\left(\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}})\right)^2>M\right\}} \\
&+2\sum_{\boldsymbol{e}\in\mathcal{E}_1}\mathbb{E}\left[\sup_{f\in\mathcal{F}_\infty}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\in\boldsymbol{n}}\varepsilon_{\boldsymbol{i}\odot\boldsymbol{e}}\left(\widetilde{f}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2\mathbb{1}_{\left\{\left(\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2\leq M\right\}}\right| \left|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\geq\mathbf{1}}\right.\right].
\end{aligned}$$

If $\overline{N^*_2} = 0$,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_\infty}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\varepsilon_{\boldsymbol{i}\odot\boldsymbol{e}}\left(\widetilde{f}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2\mathbb{1}_{\left\{\left(\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2\leq M\right\}}\right| \left|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\geq\mathbf{1}}, (\boldsymbol{i}^*)_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\right.\right]$$

is null. Otherwise $\overline{N^*_2} > 0$ and conditional on $\left((\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\geq\mathbf{1}}, (\boldsymbol{i}^*)_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\right)$, we can consider for every $\eta_1 > 0$ and $\boldsymbol{e}\in\mathcal{E}_1$ a minimal η_1-covering of $\widetilde{\mathcal{F}}^2_\infty = \{g = (\widetilde{f}_1 - \widetilde{f}_2)^2 : (\widetilde{f}_1, \widetilde{f}_2)\in\mathcal{F}\times\mathcal{F}\}$ for the seminorm

$$||g||^*_{\boldsymbol{e},M,1} = \frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\boldsymbol{e}\leq\boldsymbol{d}\leq\boldsymbol{n}\odot\boldsymbol{e}}\left|\sum_{\mathbf{1}-\boldsymbol{e}\leq\boldsymbol{d}'\leq\boldsymbol{n}\odot(\mathbf{1}-\boldsymbol{e})} g(\widetilde{Y}_{(\boldsymbol{d}+\boldsymbol{d}')^*})\right|$$

with balls centered in $\mathcal{F}$. This implies

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}_\infty}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\varepsilon_{\boldsymbol{i}\odot\boldsymbol{e}}\left(\widetilde{f}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2\mathbb{1}_{\left\{\left(\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2\leq M\right\}}\right| \left|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\geq\mathbf{1}}, (\boldsymbol{i}^*)_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\right.\right] \\
\leq& 4\sqrt{2\log 2N\left(\eta_1, \widetilde{\mathcal{F}}^2_\infty, ||.||^*_{M,1}\right)}M\frac{1}{\sqrt{\underline{n}}} + \eta_1.
\end{aligned}$$

Remark that for $\widetilde{f} \in \widetilde{\mathcal{F}}_\infty$ with corresponding $f \in \mathcal{F}_\infty$, $||\widetilde{f}^2||^*_{\boldsymbol{e},M,1} \leq \overline{N_2^*}||f^2||_{\mathbb{Q}^*_{\boldsymbol{n}},1}$ where $||g||_{\mathbb{Q}^*_{\boldsymbol{n}},1} = \frac{1}{\sum_{\mathbf{1}\leq \boldsymbol{i}\leq \boldsymbol{n}} N_{\boldsymbol{i}^*}^2} \sum_{\mathbf{1}\leq \boldsymbol{i}\leq \boldsymbol{n}} N_{\boldsymbol{i}^*} \sum_{\ell=1}^{N_{\boldsymbol{i}^*}} |g(Y_{\boldsymbol{i}^*,\ell})|$. Then, for every $\eta > 0$, using Points 1, 2 and 4 of Lemma S4.11 and letting $\eta_1 = 8\eta\overline{N_2^*}||F^2||_{\mathbb{Q}^*_{\boldsymbol{n}},1}$, we obtain

$$\sum_{\boldsymbol{e}\in\mathcal{E}_1} \mathbb{E}\left[\sup_{f\in\mathcal{F}_\infty}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq \boldsymbol{i}\leq \boldsymbol{n}} \varepsilon_{\boldsymbol{i}\odot\boldsymbol{e}}\left(\widetilde{f}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2 \mathbb{1}_{\left\{\left(\tilde{F}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2\leq M\right\}}\right| \left| (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\geq\mathbf{1}}, (\boldsymbol{i}^*)_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\right]\right.$$
$$\leq 4k\sqrt{2\log 2 \sup_Q N^2\left(\eta||F||_{Q,2}, \mathcal{F}, ||.||_{Q,2}\right)} M\frac{1}{\sqrt{\underline{n}}} + 8k\eta\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq \boldsymbol{i}\leq \boldsymbol{n}} N_{\boldsymbol{i}^*}\sum_{\ell=1}^{N_{\boldsymbol{i}^*}} F^2(Y_{\boldsymbol{i}^*,\ell}).$$

Integration with respect to $(\boldsymbol{i}^*)_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}|(\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\geq\mathbf{1}}$ leads to

$$\sum_{\boldsymbol{e}\in\mathcal{E}_1} \mathbb{E}\left[\sup_{f\in\mathcal{F}_\infty}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq \boldsymbol{i}\leq \boldsymbol{n}} \varepsilon_{\boldsymbol{i}\odot\boldsymbol{e}}\left(\widetilde{f}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2 \mathbb{1}_{\left\{\left(\tilde{F}(\widetilde{Y}_{\boldsymbol{i}^*})\right)^2\leq M\right\}}\right| \left| (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\geq\mathbf{1}}, (\boldsymbol{i}^*)_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\right]\right.$$
$$\leq 4k\sqrt{2\log 2 \sup_Q N^2\left(\eta||F||_{Q,2}, \mathcal{F}, ||.||_{Q,2}\right)} M\frac{1}{\sqrt{\underline{n}}} + 8k\eta\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq \boldsymbol{i}\leq \boldsymbol{n}} N_{\boldsymbol{i}}\sum_{\ell=1}^{N_{\boldsymbol{i}}} F^2(Y_{\boldsymbol{i},\ell}).$$

Observe that $\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\in\boldsymbol{n}}\left(\widetilde{F}(\widetilde{Y}_{\boldsymbol{i}})\right)^2 \mathbb{1}_{\left\{\left(\tilde{F}(\widetilde{Y}_{\boldsymbol{i}})\right)^2>M\right\}}$ and $\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} N_{\boldsymbol{i}}\sum_{\ell=1}^{N_{\boldsymbol{i}}} F^2(Y_{\boldsymbol{i},\ell})$ converge a.s. to $\mathbb{E}\left(\left(\widetilde{F}(\widetilde{Y}_{\mathbf{1}})\right)^2 \mathbb{1}_{\left\{\left(\tilde{F}(\widetilde{Y}_{\mathbf{1}})\right)^2>M\right\}}\right)$ and $\mathbb{E}\left(N_{\mathbf{1}}\sum_{\ell=1}^{N_{\mathbf{1}}} F^2(Y_{\mathbf{1},\ell})\right)$, by application of Point 1 of Theorem 4.6 for a class $\mathcal{F}$ reduced to a singleton. Choosing M and η arbitrarily small, we obtain, as $m\to\infty$,

$$\mathbb{E}\left(\sup_{\widetilde{f}\in\widetilde{\mathcal{F}}_\infty}\left|\mathbb{P}^*_{\boldsymbol{n}}\widetilde{f}^2 - \mathbb{P}_{\boldsymbol{n}}\widetilde{f}^2\right| \mid (\widetilde{Y}_{\boldsymbol{i}})_{\boldsymbol{i}\geq\mathbf{1}}\right) \xrightarrow{a.s.} 0.$$

4.7.3 Technical lemmas

4.7.3.1 Results related to the symmetrisation lemma

Below, Φ denotes a non-decreasing convex function Φ from $\mathbb{R}^+$ to $\mathbb{R}$.

Lemma 4.4 (A useful inequality). *Let $m \in \mathbb{N}^+$ and $(X_1, ..., X_m)$ be any random variables with values in $\mathcal{X}$ and $\mathcal{H}$ be a pointwise measurable class of functions from $\mathcal{X}$ to $\mathbb{R}$. Then*

$$\mathbb{E}\Phi\left[\sup_{h\in\mathcal{H}}\left|\sum_{j=1}^m h(X_j)\right|\right] \leq \frac{1}{m}\sum_{j=1}^m \mathbb{E}\Phi\left[m\sup_{h\in\mathcal{H}}|h(X_j)|\right].$$

Lemma 4.5 (Symmetrization, separately exchangeable, unbalanced and dissociated arrays)**.** *Let $k\in\mathbb{N}^+$, $\boldsymbol{n} = (n_1, ..., n_k) \in \mathbb{N}^{+k}$ and $\left(\widetilde{Y}_{\boldsymbol{i}}\right)_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}$ a family of random variables with values in a Polish space, such that*

$$\left(\widetilde{Y}_{\boldsymbol{i}}\right)_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} \overset{a.s.}{=} \left(\tau\left((U_{\boldsymbol{i}\odot\boldsymbol{e}})_{\boldsymbol{e}\in\cup_{r=1}^k\mathcal{E}_r}\right)\right)_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}$$

for $(U_A)_{A\in\mathbb{N}^k}$ a family of i.i.d. real random variables and some measurable function τ. Let $\mathcal{G}$ a pointwise measurable class of integrable functions of $\widetilde{Y}_{\mathbf{1}}$. We have

$$\mathbb{E}\left[\Phi\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} g\left(\widetilde{Y}_{\boldsymbol{i}}\right) - \mathbb{E}\left[g\left(\widetilde{Y}_{\mathbf{1}}\right)\right]\right|\right)\right]$$
$$\leq \frac{1}{2^k-1}\sum_{\boldsymbol{e}\in\cup_{r=1}^k\mathcal{E}_r}\mathbb{E}\left[\Phi\left(2(2^k-1)\sup_{g\in\mathcal{G}}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\varepsilon_{\boldsymbol{i}\odot\boldsymbol{e}} g\left(\widetilde{Y}_{\boldsymbol{i}}\right)\right|\right)\right],$$

with $(\varepsilon_A)_{A\in\mathbb{N}^k}$ are i.i.d. Rademacher variables, independent of $\left(\widetilde{Y}_{\boldsymbol{i}}\right)_{\boldsymbol{i}\in\mathbb{N}^{+k}}$.

4.7.3.1.1 Proof of Lemma S4.4

By the triangle inequality and properties of the supremum,

$$\sup_{h\in\mathcal{H}}\left|\sum_{j=1}^{m}h(X_j)\right|\leq\frac{1}{m}\sum_{j=1}^{m}m\sup_{h\in\mathcal{H}}\left|h(X_j)\right|.$$

The result follows by monotonicity and convexity of Φ.

4.7.3.1.2 Proof of Lemma S4.5

The proof is much simpler than that of Lemma 4.2 because there is much more invariance in separately exchangeable arrays than in jointly exchangeable ones. Consequently the decoupling and recoupling steps used in the proof of Lemma 4.2 are not necessary.

To get the result, we introduce $\left(U_A^{(1)}\right)_{A\in\mathbb{N}^k}$ which is an independent copy of $(U_A)_{A\in\mathbb{N}^k}$. We assume without loss of generality that the last argument of τ is $U_{\boldsymbol{i}\odot\mathbf{1}}=U_{\boldsymbol{i}}$. On the set $\cup_{l=1}^k\mathcal{E}_l$, $\prec$ is the strict total order used (implicitly) to enumerate the arguments of τ in the statement of the Lemma. We extend this order to $\cup_{l=0}^k\mathcal{E}_l$ considering that $\mathbf{0}\prec\boldsymbol{e}\preceq 1$ for every $\boldsymbol{e}\in\cup_{l=1}^k\mathcal{E}_l$. For every $(\boldsymbol{e},\boldsymbol{e}')\in\left(\cup_{l=0}^k\mathcal{E}_l\right)^2$, we write $\boldsymbol{e}\preceq\boldsymbol{e}'$ if $\boldsymbol{e}\prec\boldsymbol{e}'$ or $\boldsymbol{e}=\boldsymbol{e}'$. We also let $\widetilde{Y}_{\boldsymbol{i}}^{(\boldsymbol{e})}=\tau\left(\left(U_{\boldsymbol{i}\odot\boldsymbol{e}}^{(1)}\right)_{\mathbf{0}\prec\boldsymbol{e}'\preceq\boldsymbol{e}},(U_{\boldsymbol{i}\odot\boldsymbol{e}'})_{\boldsymbol{e}\prec\boldsymbol{e}'\preceq\mathbf{1}}\right)$ for every $\boldsymbol{e}\in\cup_{l=1}^k\mathcal{E}_l$ (hence $\widetilde{Y}_{\boldsymbol{i}}=\widetilde{Y}_{\boldsymbol{i}}^{(\mathbf{0})}$). Convexity of Φ then implies

$$\begin{aligned}
&\mathbb{E}\left[\Phi\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}g\left(\widetilde{Y}_{\boldsymbol{i}}\right)-\mathbb{E}\left[g\left(\widetilde{Y}_{\mathbf{1}}\right)\right]\right|\right)\right]\\
\leq&\mathbb{E}\left[\Phi\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}g\left(\widetilde{Y}_{\boldsymbol{i}}^{(\mathbf{0})}\right)-g\left(\widetilde{Y}_{\boldsymbol{i}}^{(\mathbf{1})}\right)\right|\right)\right]\\
=&\mathbb{E}\left[\Phi\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\sum_{\mathbf{0}\prec\boldsymbol{e}\preceq\mathbf{1}}g\left(\widetilde{Y}_{\boldsymbol{i}}^{(\boldsymbol{e}_{prec})}\right)-g\left(\widetilde{Y}_{\boldsymbol{i}}^{(\boldsymbol{e})}\right)\right|\right)\right]\\
\leq&\frac{1}{2^k-1}\sum_{\mathbf{0}\prec\boldsymbol{e}\preceq\mathbf{1}}\mathbb{E}\left[\Phi\left((2^k-1)\sup_{g\in\mathcal{G}}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}g\left(\widetilde{Y}_{\boldsymbol{i}}^{(\boldsymbol{e}_{prec})}\right)-g\left(\widetilde{Y}_{\boldsymbol{i}}^{(\boldsymbol{e})}\right)\right|\right)\right]\\
=&\frac{1}{2^k-1}\sum_{\boldsymbol{e}\in\cup_{l=1}^k\mathcal{E}_l}\mathbb{E}\left[\Phi\left((2^k-1)\sup_{g\in\mathcal{G}}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}g\left(\widetilde{Y}_{\boldsymbol{i}}^{(\boldsymbol{e}_{prec})}\right)-g\left(\widetilde{Y}_{\boldsymbol{i}}^{(\boldsymbol{e})}\right)\right|\right)\right],
\end{aligned}$$

with $\boldsymbol{e}_{prec}$ the element that precedes $\boldsymbol{e}$ for the strict total order $\prec$. For every $\boldsymbol{e}\in\cup_{l=1}^k\mathcal{E}_l$, note that

$$\begin{aligned}
&\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}g\left(\widetilde{Y}_{\boldsymbol{i}}^{(\boldsymbol{e}_{prec})}\right)-g\left(\widetilde{Y}_{\boldsymbol{i}}^{(\boldsymbol{e})}\right)\\
=&\sum_{\boldsymbol{e}\leq\boldsymbol{d}\leq\boldsymbol{n}\odot\boldsymbol{e}}\ \sum_{\mathbf{1}-\boldsymbol{e}\leq\boldsymbol{d}'\leq\boldsymbol{n}\odot(\mathbf{1}-\boldsymbol{e})}g\left(\widetilde{Y}_{\boldsymbol{d}+\boldsymbol{d}'}^{(\boldsymbol{e}_{prec})}\right)-g\left(\widetilde{Y}_{\boldsymbol{d}+\boldsymbol{d}'}^{(\boldsymbol{e})}\right).
\end{aligned}$$

Furthermore,

$$\left(\sum_{\mathbf{1}-\boldsymbol{e}\leq\boldsymbol{d}'\leq\boldsymbol{n}\odot(\mathbf{1}-\boldsymbol{e})}g\left(\widetilde{Y}_{\boldsymbol{d}+\boldsymbol{d}'}^{(\boldsymbol{e}_{prec})}\right)-g\left(\widetilde{Y}_{\boldsymbol{d}+\boldsymbol{d}'}^{(\boldsymbol{e})}\right)\right)_{\boldsymbol{e}\leq\boldsymbol{d}\leq n\odot\boldsymbol{e}}$$

is an array of independent and symmetric random variables conditional on $\left(\left(U_{\boldsymbol{i}\odot\boldsymbol{e}}^{(1)}\right)_{\mathbf{0}\prec\boldsymbol{e}'\prec\boldsymbol{e}},(U_{\boldsymbol{i}\odot\boldsymbol{e}})_{\boldsymbol{e}\prec\boldsymbol{e}'\preceq\mathbf{1}}\right)$.

Standard symmetrization arguments [see for instance 137, Lemma 2.3.1 in the i.i.d. case] entail

$$\begin{aligned}
&\mathbb{E}\left[\Phi\left((2^k-1)\sup_{g\in\mathcal{G}}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\boldsymbol{e}\leq\boldsymbol{d}\leq\boldsymbol{n}\odot\boldsymbol{e}}\ \sum_{\mathbf{1}-\boldsymbol{e}\leq\boldsymbol{d}'\leq\boldsymbol{n}\odot(\mathbf{1}-\boldsymbol{e})} g\left(\widetilde{Y}^{(\boldsymbol{e}_{prec})}_{\boldsymbol{d}+\boldsymbol{d}'}\right)-g\left(\widetilde{Y}^{(\boldsymbol{e})}_{\boldsymbol{d}+\boldsymbol{d}'}\right)\right|\right)\right]\\
\leq&\mathbb{E}\left[\Phi\left(2(2^k-1)\sup_{g\in\mathcal{G}}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\boldsymbol{e}\leq\boldsymbol{d}\leq\boldsymbol{n}\odot\boldsymbol{e}}\varepsilon_{\boldsymbol{d}}\sum_{\mathbf{1}-\boldsymbol{e}\leq\boldsymbol{d}'\leq\boldsymbol{n}\odot(\mathbf{1}-\boldsymbol{e})} g\left(\widetilde{Y}_{\boldsymbol{d}+\boldsymbol{d}'}\right)\right|\right)\right]\\
=&\mathbb{E}\left[\Phi\left(2(2^k-1)\sup_{g\in\mathcal{G}}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\varepsilon_{\boldsymbol{i}\odot\boldsymbol{e}}g\left(\widetilde{Y}_{\boldsymbol{i}}\right)\right|\right)\right].
\end{aligned}$$

4.7.3.2 Results related to laws of large numbers

Lemma 4.6. *Under Assumptions 4.1-4.3,* $\mathbb{E}\left(\sup_{\mathcal{F}}|\mathbb{P}_n^*f-\mathbb{P}_nf|\,\middle|\,(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)$ *tends to 0 almost surely as* $n\to\infty$*.*

Lemma 4.7. *Suppose that Assumptions 4.2, 4.3 and 4.5 hold and* $\boldsymbol{n}=(n_1(m),...,n_k(m))$ *satisfies* $n_j(m)\to+\infty$ *as* $m\to+\infty$ *for every* $j=1,...,k$*. Then* $\mathbb{E}\left(\sup_{\mathcal{F}}\left|\widetilde{\mathbb{P}}^*_{\boldsymbol{n}}f-\widetilde{\mathbb{P}}_{\boldsymbol{n}}f\right|\,\middle|\,(N_{\boldsymbol{i}},(Y_{\boldsymbol{i},\ell})_{N_{\boldsymbol{i}}\geq\ell\geq1})_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\right)$ *tends to 0 almost surely as* $m\to\infty$*.*

Lemma 4.8 (Control of sums of quadratic terms)**.**
If Assumption 4.1 holds and $\mathbb{E}\left[Y_{\mathbf{1}}^2\right]<+\infty$*, then for* $h(\boldsymbol{i})=\mathbb{1}_{\{\boldsymbol{i}\in\mathbb{I}_{n,k}\}}\sum_{\pi\in\mathfrak{S}_k}Y_{\boldsymbol{j}_\pi}$ *we have for every* $j=0,...,k$

$$\begin{aligned}
&\sum_{\boldsymbol{i}\in\{1,...,n\}^{2k-j}}h(i_1,...,i_k)h(i_1,...,i_l,i_{k+1},...,i_{2k-j})\\
=&\sum_{c=0}^{k-j}\binom{k-j}{c}^2\left(n^{2k-j-c}\mathbb{E}\left[h(1,...,k)h(1,...,j+c,k+1,...,2k-c-j)\right]+o_{a.s}(n^{2k-j-c})\right).
\end{aligned}$$

Lemma 4.9 (Control of sums of quadratic terms under separate exchangeability)**.**
Let $h(\boldsymbol{i})=\sum_{\ell=1}^{N_{\boldsymbol{i}}}Y_{\boldsymbol{i},\ell}$*. Suppose Assumption 4.5 holds,* $\mathbb{E}\left[Y_{\mathbf{1}}^2\right]<+\infty$ *and* $\boldsymbol{n}=(n_1(m),...,n_k(m))\in\mathbb{N}^{+k}$ *satisfies* $n_j(m)\to+\infty$ *when* $m\to+\infty$ *for every* $j=1,...,k$*. Then for every* $\boldsymbol{e}\in\cup_{r=1}^k\mathcal{E}_r$

$$\frac{1}{\prod_{r=1}^k n_r\mathbb{1}_{\{e_r=1\}}}\frac{1}{\prod_{r=1}^k n_r(n_r-1)\mathbb{1}_{\{e_r=0\}}}\sum_{(\boldsymbol{i},\boldsymbol{i}')\in\mathcal{I}_{\boldsymbol{n},\boldsymbol{e}}}h(\boldsymbol{i})h(\boldsymbol{i}')=\mathbb{E}\left[h(\mathbf{1})h(\boldsymbol{b}_{\boldsymbol{e}})\right]+o_{a.s}(1),$$

where $\boldsymbol{b}_{\boldsymbol{e}}$ *is a* k*-dimensional vector such that its* j*-th entry is equal to 1 if* $e_j=1$ *and 2 otherwise and* $\mathcal{I}_{\boldsymbol{n},\boldsymbol{e}}=\{(\boldsymbol{i},\boldsymbol{i}'):\mathbf{1}\leq\boldsymbol{i},\boldsymbol{i}'\leq\boldsymbol{n},\ i_r=i'_r$ *if* $e_r=1$ *and* $i_r\neq i'_r$ *otherwise*$\}$*.*

4.7.3.2.1 Proof of Lemma S4.6

Let i^* the ith index sampled with replacement in $\{1,...,n\}$. The i^*s are distributed as $i^*\overset{i.i.d}{\sim}\mathcal{U}_{\{1,...,n\}}$. For every $\boldsymbol{i}=(i_1,...,i_k)\in\mathbb{I}_{n,k}$, $\boldsymbol{i}^*$ stands for $(i_1^*,...,i_k^*)$. Conditional on the data and for every $f\in\mathcal{F}$, $\mathbb{P}_n^*f=\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}f\left(Y_{\boldsymbol{i}^*}\right)\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}$. We remark $\mathbb{E}\left(f\left(Y_{\boldsymbol{i}^*}\right)\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\mid(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right)=P_n'f=\mathbb{E}\left[\mathbb{P}_n^*f\mid(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right]$.

Note that conditionally on $(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}$, $\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}f(Y_{\boldsymbol{i}^*})\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}$ is a U-statistics since $f(Y_{\boldsymbol{i}^*})\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}$ admits a representation $f(\tau(U_{i_1},...,U_{i_k}))\mathbb{1}_{\left\{(U_{i_1},...,U_{i_k})\in\mathbb{I}_{n,k}\right\}}$ for i.i.d. $U_i=i^*$. We also have that $\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}f(Y_{\boldsymbol{i}^*})\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}=\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}h(\boldsymbol{i}^*)$ with $h:\boldsymbol{i}\mapsto\frac{1}{k!}\sum_{\pi\in\mathfrak{S}_k}f(Y_{\boldsymbol{i}_\pi})\mathbb{1}_{\{\boldsymbol{i}_\pi\in\mathbb{I}_{n,k}\}}$. As a result, the inequality proved on page 1508 in [9] is valid with their f replaced with h (in particular, the sixth inequality on the latter page is true as h is symmetric in its arguments and $h(\cdot)$ does not depend on

$\boldsymbol{i}$) and we can write for some constant C_k that depends on k only

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\left|\mathbb{P}_n^* f-\mathbb{P}_n' f\right|\middle|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right]\leq kC_k\mathbb{E}\left[\sup_{f\in\mathcal{F}_\delta}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{i_1\}}f\left(Y_{\boldsymbol{i}^*}\right)\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right|\middle|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right].$$

Let $N^*=\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}$. If $N^*=0$, we sample fewer than k different units in the bootstrap. In that case, the supremum of the Rademacher process is always equal to 0. As a result,

$$\begin{aligned}&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{i_1\}}f\left(Y_{\boldsymbol{i}^*}\right)\mathbb{1}_{\{F(Y_{\boldsymbol{i}^*})\leq M\}}\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right|\middle|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right]\\ =&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{i_1\}}f\left(Y_{\boldsymbol{i}^*}\right)\mathbb{1}_{\{F(Y_{\boldsymbol{i}^*})\leq M\}}\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right|\middle|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k},N^*>0\right]\mathbb{P}\left(N^*>0\right).\end{aligned}$$

We now adapt the steps of the proof of Theorem 4.1. Conditional on $\left((Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k},(\boldsymbol{i}^*)_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\right)$ and $N^*>0$, we can consider for every $\eta_1>0$ and every $\boldsymbol{e}\in\mathcal{E}_1$ a minimal η_1-covering of $\mathcal{F}$ for the seminorm

$$||g||_{M,1}^*=\frac{(n-k)!}{n!}\sum_{i_1=1}^n\left|\sum_{(i_2,\ldots,i_k):\boldsymbol{i}\in\mathbb{I}_{n,k}}g\left(Y_{\boldsymbol{i}^*}\right)\mathbb{1}_{\{F()\leq M\}}\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right|$$

with balls centered in $\mathcal{F}$. This implies

$$\begin{aligned}&\mathbb{E}\left[\sup_{\mathcal{F}}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{\boldsymbol{i}\odot\boldsymbol{e}\}^+}f\left(Y_{\boldsymbol{i}^*}\right)\mathbb{1}_{\{F(Y_{\boldsymbol{i}^*})\leq M\}}\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right|\middle|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k},(\boldsymbol{i}^*)_{\boldsymbol{i}\in\mathbb{I}_{n,k}},N^*>0\right]\\ \leq&\sqrt{2\log 2N\left(\eta_1,\mathcal{F},||.||_{M,1}^*\right)}M\frac{1}{\sqrt{n}}+\eta_1.\end{aligned}$$

Remark that $||g||_{M,1}^*\leq N^*||g||_{\mathbb{Q}_n,1}^*$ where $||g||_{\mathbb{Q}_n,1}^*=N^{*-1}\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}|g(Y_{\boldsymbol{i}^*})|\,\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}$, for $\mathbb{Q}_n=N^{*-1}\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\delta_{\{Y_{\boldsymbol{i}^*}\}}\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}$ a (random) probability measure with finite support on $\mathcal{Y}$ that is well-defined when $N^*>0$. Then, for every $\eta>0$, letting $\eta_1=\eta N^*||F||_{\mathbb{Q}_n,1}^*$ and using Point 2 of Lemma S4.11 and Point 1 of Lemma S4.11,

$$\begin{aligned}&\mathbb{E}\left[\sup_{\mathcal{F}}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{i_1\}}f\left(Y_{\boldsymbol{i}^*}\right)\mathbb{1}_{\{F(Y_{\boldsymbol{i}^*})\leq M\}}\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right|\middle|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k},(\boldsymbol{i}^*)_{\boldsymbol{i}\in\mathbb{I}_{n,k}},N^*>0\right]\\ \leq&\sqrt{2\log 2\sup_Q N\left(\eta||F||_{Q,1},\mathcal{F},||.||_{Q,1}\right)}M\frac{1}{\sqrt{n}}+\eta N^*||F||_{\mathbb{Q}_n,1}^*.\end{aligned}$$

Integration with respect to $(\boldsymbol{i}^*)_{\boldsymbol{i}\in\mathbb{I}_{n,k}}|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k},N^*>0$ combined with the fact that $\mathbb{E}\left[N^*||F||_{\mathbb{Q}_n,1}^*\mid(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k},N^*>0\right]=\mathbb{E}\left[N^*||F||_{\mathbb{Q}_n,1}^*\mid(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right]/\mathbb{P}\left(N^*>0\right)$ leads to

$$\begin{aligned}&\mathbb{E}\left[\sup_{\mathcal{F}}\left|\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\varepsilon_{\{i_1\}}f\left(Y_{\boldsymbol{i}^*}\right)\mathbb{1}_{\{F(Y_{\boldsymbol{i}^*})\leq M\}}\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\right|\middle|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right]\\ \leq&\sqrt{2\log 2\sup_Q N\left(\eta||F||_{Q,1},\mathcal{F},||.||_{Q,1}\right)}M\frac{1}{\sqrt{n}}+\eta\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}\mathbb{E}\left[F(Y_{\boldsymbol{i}^*})\mathbb{1}_{\{\boldsymbol{i}^*\in\mathbb{I}_{n,k}\}}\middle|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right]\\ =&\sqrt{2\log 2\sup_Q N\left(\eta||F||_{Q,1},\mathcal{F},||.||_{Q,1}\right)}M\frac{1}{\sqrt{n}}+\eta\frac{1}{n^k}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}F(Y_{\boldsymbol{i}}).\end{aligned}$$

We observe $\frac{1}{n^k}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}F(Y_{\boldsymbol{i}})\mathbb{1}_{\{Y_{\boldsymbol{i}}>M\}}$ and $\frac{1}{n^k}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}F(Y_{\boldsymbol{i}})$ converge a.s. to $\mathbb{E}\left(F(Y_{\mathbf{1}})\mathbb{1}_{\{Y_{\mathbf{1}}>M\}}\right)$ and $\mathbb{E}\left(F(Y_{\mathbf{1}})\right)$ by almost sure convergence of the sample mean of jointly exchangeable arrays [66] and

ergodicity of dissociated arrays [96] or Theorem 4.1 for a class $\mathcal{F}$ reduced to a singleton. Choosing M and η such that $\mathbb{E}\left(F(Y_{\mathbf{1}})\mathbb{1}_{\{Y_{\mathbf{1}}>M\}}\right)+\eta\mathbb{E}\left(F(Y_{\mathbf{1}})\right)$ is arbitrarily small, we deduce that for $n\to\infty$

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{P}_n^*f-\mathbb{P}_n^{'}f\right|\left|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right.\right]\xrightarrow{a.s.}0.$$

Finally, the triangle inequality enables us to write

$$\begin{aligned}&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{P}_n^*f-\mathbb{P}_nf\right|\left|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right.\right]\\ \leq&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\left(\frac{(n-k)!}{n!}-\frac{1}{n^k}\right)\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}f(Y_{\boldsymbol{i}})\right|\left|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right.\right]+\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{P}_n^*f-\mathbb{P}_n^{'}f\right|\left|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right.\right]\\ \leq&\left(1-\frac{n!}{n^k(n-k)!}\right)\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}F(Y_{\boldsymbol{i}})+\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{P}_n^*f-\mathbb{P}_n^{'}f\right|\left|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right.\right].\end{aligned}$$

Because $\frac{(n-k)!}{n!}\sum_{\boldsymbol{i}\in\mathbb{I}_{n,k}}F(Y_{\boldsymbol{i}})\xrightarrow{a.s.}\mathbb{E}(F(Y_{\mathbf{1}}))$ and $\frac{n!}{n^k(n-k)!}\to1$, we conclude

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{P}_n^*f-\mathbb{P}_nf\right|\left|(Y_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathbb{I}_k}\right.\right]\xrightarrow{a.s.}0.$$

4.7.3.2.2 Proof of Lemma S4.7

For every $j=1,...,k$, let i_j^* the i_j-th index sampled with replacement in $[1;n_j]$. The i_j^*s are distributed as $i_j^*\overset{i.i.d}{\sim}\mathcal{U}_{[1;n_j]}$ and the k sequences $(i_1^*)_{i_1=1}^{n_1},...,(i_k^*)_{i_k=1}^{n_k}$ are also mutually independent. For every $\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}$, $\boldsymbol{i}^*$ denotes $(i_1^*,...,i_k^*)$. Conditional on the data and for every $f\in\mathcal{F}$, $\widetilde{\mathbb{P}}_{\boldsymbol{n}}^*f=\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\widetilde{f}\,(\boldsymbol{i}^*)$ with $\widetilde{f}\,(\boldsymbol{i}^*)=\sum_{\ell=1}^{N_{\boldsymbol{i}^*}}f\left(Y_{\boldsymbol{i}^*,\ell}\right)$. We have: $\mathbb{E}\left[\mathbb{P}_n^*f\mid(N_{\boldsymbol{i}},(Y_{\boldsymbol{i},\ell})_{\ell\geq1})_{\boldsymbol{i}\in\mathbb{N}^{+k}}\right]=\widetilde{\mathbb{P}}_{\boldsymbol{n}}f$. Note that conditional on $(N_{\boldsymbol{i}},(Y_{\boldsymbol{i},\ell})_{\ell\geq1})_{\boldsymbol{i}\in\mathbb{N}^{+k}}$, $(\boldsymbol{i}^*)_{\boldsymbol{i}\in\mathbb{I}_{n,k}}$ is a family of random vectors that admit a representation $\boldsymbol{i}^*=\tau((U_{\boldsymbol{i}\odot\boldsymbol{e}})_{\boldsymbol{e}\in\mathcal{E}_1})$ with $(U_{\boldsymbol{i}})_{\mathbf{0}\leq\boldsymbol{i}\leq\boldsymbol{n}}$ i.i.d. random variables (consider $\tau:(u_1,...,u_k)\in[0,1]^k\mapsto(\lceil n_1\times u_1\rceil,...,\lceil n_k\times u_k\rceil)$ where $\lceil\cdot\rceil$ denotes the ceiling function and $U_{\boldsymbol{i}}\sim\mathcal{U}_{[0,1]}$). As a result, conditionally on the data, Lemma S4.5 applies to $\widetilde{Y}_{\boldsymbol{i}}=\boldsymbol{i}^*$, $\mathcal{G}=\left\{\widetilde{f}:\widetilde{f}\,(\boldsymbol{i}^*)=\sum_{\ell=1}^{N_{\boldsymbol{i}^*}}f\left(Y_{\boldsymbol{i}^*,\ell}\right),f\in\mathcal{F}\right\}$ and $\Phi=\mathrm{Id}$. Moreover, because only terms involving $\boldsymbol{e}\in\mathcal{E}_1$ appear in the representation of $\boldsymbol{i}^*$, a simplification of the proof of Lemma S4.5 leads to the following inequality

$$\begin{aligned}&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\widetilde{\mathbb{P}}_{\boldsymbol{n}}^*f-\widetilde{P}_{\boldsymbol{n}}f\right|\left|(N_{\boldsymbol{i}},(Y_{\boldsymbol{i},\ell})_{\ell\geq1})_{\boldsymbol{i}\in\mathbb{N}^{+k}}\right.\right]\\ \leq&\frac{2}{\Pi_{\boldsymbol{n}}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}\sum_{\ell=1}^{N_{\boldsymbol{i}}}F(Y_{\boldsymbol{i},\ell})\mathbb{1}_{\{\sum_{\ell=1}^{N_{\boldsymbol{i}}}F(Y_{\boldsymbol{i},\ell})>M\}}\\ &+2\sum_{\boldsymbol{e}\in\mathcal{E}_1}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\varepsilon_{\boldsymbol{i}\odot\boldsymbol{e}}\sum_{\ell=1}^{N_{\boldsymbol{i}^*}}f\left(Y_{\boldsymbol{i}^*,\ell}\right)\mathbb{1}_{\{\sum_{\ell=1}^{N_{\boldsymbol{i}^*}}F(Y_{\boldsymbol{i}^*,\ell})\leq M\}}\right|\left|(N_{\boldsymbol{i}},(Y_{\boldsymbol{i},\ell})_{\ell\geq1})_{\boldsymbol{i}\in\mathbb{N}^{+k}}\right.\right].\end{aligned}\tag{4.34}$$

The rest of the proof is similar to that of $\sup_{\mathcal{F}}\left|\widetilde{\mathbb{P}}_{\boldsymbol{n}}f-\widetilde{P}f\right|\xrightarrow{L^1}0$: in fact, with $||\widetilde{f}||_{\boldsymbol{e},M,1}$ redefined as $||\widetilde{f}||_{\boldsymbol{e},M,1}=\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\boldsymbol{e}\leq\boldsymbol{d}\leq\boldsymbol{n}\odot\boldsymbol{e}}\left|\sum_{\mathbf{1}-\boldsymbol{e}\leq d'\leq n\odot(\mathbf{1}-\boldsymbol{e})}\sum_{\ell=1}^{N_{\boldsymbol{i}^*}}f\left(Y_{\boldsymbol{i}^*,\ell}\right)\mathbb{1}_{\left\{\sum_{\ell=1}^{N_{\boldsymbol{i}^*}}F\left(Y_{\boldsymbol{i}^*,\ell}\right)\leq M\right\}}\right|$, we have for every

$\boldsymbol{e} \in \mathcal{E}_1$, $M > 0$ and $\eta_1 \geq 0$ (possibly random)

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\varepsilon_{\boldsymbol{i}\odot\boldsymbol{e}}\sum_{\ell=1}^{N_{\boldsymbol{i}^*}} f\left(Y_{\boldsymbol{i}^*,\ell}\right)\mathbb{1}_{\{\sum_{\ell=1}^{N_{\boldsymbol{i}^*}} F(Y_{\boldsymbol{i}^*,\ell})\leq M\}}\right|\middle|(N_{\boldsymbol{i}},(Y_{\boldsymbol{i},\ell})_{\ell\geq1})_{\boldsymbol{i}\in\mathbb{N}^{+k}}\right]\\
\leq&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}}\varepsilon_{\boldsymbol{i}\odot\boldsymbol{e}}\sum_{\ell=1}^{N_{\boldsymbol{i}^*}} f\left(Y_{\boldsymbol{i}^*,\ell}\right)\mathbb{1}_{\{\sum_{\ell=1}^{N_{\boldsymbol{i}^*}} F(Y_{\boldsymbol{i}^*,\ell})\leq M\}}\right|\middle|(N_{\boldsymbol{i}},(Y_{\boldsymbol{i},\ell})_{\ell\geq1})_{\boldsymbol{i}\in\mathbb{N}^{+k}},\overline{N_1^*}>0\right]\\
&\times\mathbb{P}\left(\overline{N_1^*}>0\middle|(N_{\boldsymbol{i}},(Y_{\boldsymbol{i},\ell})_{\ell\geq1})_{\boldsymbol{i}\in\mathbb{N}^{+k}}\right)\\
\leq&\mathbb{E}\left[\sqrt{2\log 2N\left(\eta_1,\mathcal{F},||.||_{\boldsymbol{e},M,1}\right)}M\frac{1}{\sqrt{\underline{n}}}+\eta_1\middle|(N_{\boldsymbol{i}},(Y_{\boldsymbol{i},\ell})_{\ell\geq1})_{\boldsymbol{i}\in\mathbb{N}^{+k}},\overline{N_1^*}>0\right]\\
&\times\mathbb{P}\left(\overline{N_1^*}>0\middle|(N_{\boldsymbol{i}},(Y_{\boldsymbol{i},\ell})_{\ell\geq1})_{\boldsymbol{i}\in\mathbb{N}^{+k}}\right),
\end{aligned}$$

with $\overline{N_1^*} = \frac{1}{\Pi_{\boldsymbol{n}}}\sum_{\mathbf{1}\leq\boldsymbol{i}\leq\boldsymbol{n}} N_{\boldsymbol{i}^*}$.

4.7.3.2.3 Proof of Lemma S4.8

By definition of $h(\cdot)$, we have

$$\begin{aligned}
&\sum_{\boldsymbol{i}\in\{1,...,n\}^{2k-j}} h(i_1,...,i_k)h(i_1,...,i_j,i_{k+1},...,i_{2k-j})\\
=&\sum_{\boldsymbol{i}\in\overline{\{1,...,n\}^j}}\ \sum_{\boldsymbol{i}'\in\overline{(\{1,...,n\}\setminus\{\boldsymbol{i}\})^{k-j}}}\ \sum_{\boldsymbol{i}''\in\overline{(\{1,...,n\}\setminus\{\boldsymbol{i}\})^{k-j}}} h(\boldsymbol{i},\boldsymbol{i}')h(\boldsymbol{i},\boldsymbol{i}'')\\
=&\sum_{c=0}^{k-j}\binom{k-j}{c}^2\sum_{\boldsymbol{i}\in\overline{\{1,...,n\}^{j+c}}}\ \sum_{\boldsymbol{i}'\in\overline{(\{1,...,n\}\setminus\{\boldsymbol{i}\})^{k-j-c}}}\ \sum_{\boldsymbol{i}''\in\overline{(\{1,...,n\}\setminus(\{\boldsymbol{i}\}\cup\{\boldsymbol{i}'\}))^{k-j-c}}} h(\boldsymbol{i},\boldsymbol{i}')h(\boldsymbol{i},\boldsymbol{i}'').
\end{aligned}$$

Since h is invariant by permutation of its entries, the last equality holds by distinguishing between cases depending on the number of common values in the vectors $(i_{j+1},...,i_k)$ and $(i_{k+1},...,i_{2k-j})$.
As $(h(\boldsymbol{i}))_{\boldsymbol{i}\in\mathbb{I}_{n,k}}$ is a k-dimensional jointly exchangeable array,

$$\left(h(\boldsymbol{i},\boldsymbol{i}')h(\boldsymbol{i},\boldsymbol{i}'')\right)_{\boldsymbol{i}\in\overline{\{1,...,n\}^{j+c}},\boldsymbol{i}'\in\overline{(\{1,...,n\}\setminus\{\boldsymbol{i}\})^{k-j-c}},\boldsymbol{i}''\in\overline{(\{1,...,n\}\setminus(\{\boldsymbol{i}\}\cup\{\boldsymbol{i}'\}))^{k-j-c}}}$$

is a $(2k-j-c)$-dimensional jointly exchangeable array. Moreover $\mathbb{E}(Y_{\mathbf{1}}^2) < +\infty$ ensures that $\mathbb{E}\left(|h(1,...,k)h(1,...,j+c,k+1,...,2k-j-c)|\right) < +\infty$ so that Theorem 4.1 can be applied to a class $\mathcal{F}$ reduced to the identity function. The equivalence $\frac{n!}{(n-(2k-j-c))!} \sim n^{2k-j-c}$ concludes the proof.

4.7.3.2.4 Proof of Lemma S4.9

Let $\widetilde{Y}_{\boldsymbol{i}}$ stand for $(N_{\boldsymbol{i}},(Y_{\boldsymbol{i},\ell})_{1\leq\ell\leq N_{\boldsymbol{i}}})$. Let $\Sigma_{m,\boldsymbol{e}}$ the σ-algebra generated by the set of functions g from $\mathcal{D}^{\mathbb{N}^{+k}}\times\mathcal{D}^{\mathbb{N}^{+k}}$ to $\mathbb{R}$ such that:

$$g((\widetilde{Y}_{\boldsymbol{i}},\widetilde{Y}_{\boldsymbol{i}'})_{(\boldsymbol{i},\boldsymbol{i}')\in\mathcal{I}_{\boldsymbol{n},\boldsymbol{e}}}) = g((\widetilde{Y}_{\pi_1(i_1),...,\pi_k(i_k)},\widetilde{Y}_{\pi_1(i'_1),...,\pi_k(i'_k)})_{(\boldsymbol{i},\boldsymbol{i}')\in\mathcal{I}_{\boldsymbol{n},\boldsymbol{e}}}),$$

for every set of permutations $\pi_1,....,\pi_k$ such that for every $r=1,...,k$, $\pi_r(i)=i$ if $i\geq n_r$. Let $W_m = \frac{1}{\prod_{r=1}^k n_r\mathbb{1}_{\{e_r=1\}}}\frac{1}{\prod_{r=j+1}^k n_r(n_r-1)\mathbb{1}_{\{e_r=0\}}}\sum_{(\boldsymbol{i},\boldsymbol{i}')\in\mathcal{I}_{\boldsymbol{n},\boldsymbol{e}}} h(\boldsymbol{i})h(\boldsymbol{i}')$. By construction, we have for every $\underline{n}\in\mathbb{N}^+$

$$W_m = \mathbb{E}\left[W_m \mid \Sigma_{m,\boldsymbol{e}}\right] = \mathbb{E}\left[h(\mathbf{1})h(\boldsymbol{b}_{\boldsymbol{e}}) \mid \Sigma_{m,\boldsymbol{e}}\right]$$

Furthermore, $\Sigma_{m,\boldsymbol{e}} \supseteq \Sigma_{m+1,\boldsymbol{e}}$ so that

$$\mathbb{E}\left[W_m \mid \Sigma_{m+1,\boldsymbol{e}}\right] = \mathbb{E}\left[\mathbb{E}\left[h(\mathbf{1})h(\boldsymbol{b}_{\boldsymbol{e}}) \mid \Sigma_{m,\boldsymbol{e}}\right] \mid \Sigma_{m+1,\boldsymbol{e}}\right] = \mathbb{E}\left[h(\mathbf{1})h(\boldsymbol{b}_{\boldsymbol{e}}) \mid \Sigma_{m+1,\boldsymbol{e}}\right] = W_{m+1}.$$

As a result, we can conclude that $(W_m, \Sigma_{m,\boldsymbol{e}})_{m\geq 1}$ is a reverse martingale. From this follows that $W_m \xrightarrow{a.s.} \mathbb{E}\left[h(\mathbf{1})h(\boldsymbol{b_e}) \mid \Sigma_{\infty,\boldsymbol{e}}\right]$ where $\Sigma_{\infty,\boldsymbol{e}} = \cap_{m\geq 1}\Sigma_{m,\boldsymbol{e}}$ (see for instance Theorem 22 of Chapter 24 in [73]). By the dissociation assumption, this sigma-algebra is trivial (see Lemma 7.35 in [96]), hence $W_m \xrightarrow{a.s.} \mathbb{E}\left[h(\mathbf{1})h(\boldsymbol{b_e})\right]$.

4.7.3.3 Covering and entropic integrals

Lemma 4.10 (Properties of entropic integrals)**.**
Let $\mathcal{F}$ a class of functions with envelope F such that $\int_0^\infty \zeta(\varepsilon)d\varepsilon < +\infty$, with

$$\zeta(\varepsilon) = \sup_Q \sqrt{\log(N(\varepsilon||F||_{Q,2}, \mathcal{F}, ||.||_{Q,2}))}.$$

1. *$u \mapsto J_{\mathcal{F}}(u) = \int_0^u \zeta(\varepsilon)d\varepsilon$ is positive, non-decreasing, concave, larger than $u\zeta(u)$ for every $u > 0$ and $\sup_{u\geq 0} J_{\mathcal{F}}(u) = J_{\mathcal{F}}(2)$.*
2. *For every $K > 0$, $(x, y) \in [0, \infty) \times (0, \infty) \mapsto \sqrt{y}J_{\mathcal{F}}\left(K\frac{\sqrt{x}}{\sqrt{y}}\right)$ is concave.*

Lemma 4.11 (Covering numbers inequalities)**.**
For every $\varepsilon > 0$:

1. *for every class $\mathcal{H}$, every norm $||.||$ and every $\lambda > 0$: $N(\varepsilon, \mathcal{H}, \lambda||.||)) = N(\varepsilon/\lambda, \mathcal{H}, ||.||))$.*
2. *for every class $\mathcal{H}$, every pair of norms $||.|| \leq ||.||'$: $N(\varepsilon, \mathcal{H}, ||.||) \leq N(\varepsilon, \mathcal{H}, ||.||')$.*
3. *for every $\mathcal{H} \subset \mathcal{H}'$ and every norm $||.||$: $N(\varepsilon, \mathcal{H}, ||.||) \leq N(\varepsilon/2, \mathcal{H}', ||.||)$.*
4. *for every norm $||.||$, every class $\mathcal{F}$ and for $\mathcal{F}_\infty = \{f : f = f_1 - f_2, (f_1, f_2) \in \mathcal{F} \times \mathcal{F}\}$: $N(\varepsilon, \mathcal{F}_\infty, ||.||) \leq N^2(\varepsilon/2, \mathcal{F}, ||.||)$.*
5. *for every class $\mathcal{F}$ and for $\mathcal{F}_\infty^2 = \{f : f = (f_1 - f_2)^2, (f_1, f_2) \in \mathcal{F} \times \mathcal{F}\}$: $\sup_Q N(8\varepsilon||F^2||_{Q,1}, \mathcal{F}_\infty^2, ||.||_{Q,1}) \leq \sup_Q N^2(\varepsilon||F||_{Q,2}, \mathcal{F}, ||.||_{Q,2})$ where the supremum is taken over the set of all finite probability measures on the domain of the functions in $\mathcal{F}$.*

4.7.3.3.1 Proof of Lemma S4.10

1. ζ is nonnegative and nonincreasing. It follows that $u \mapsto J_{\mathcal{F}}(u)$ is positive, non-decreasing and concave. Furthermore, $J_{\mathcal{F}}(u) \geq \int_0^u \zeta(u)d\varepsilon = u\zeta(u)$ for every $u > 0$. For $\varepsilon \geq 2$, we have $N(\varepsilon||F||_{Q,2}, \mathcal{F}, ||.||_{Q,2}) = 1$ for every probability measure Q. As a result, $\zeta(\varepsilon) = 0$.

2. J is concave on $[0, \infty)$ which implies for $\lambda \in (0; 1)$, $(x, x') \in [0, \infty)^2$, $(y, y') \in (0, \infty)^2$

$$\begin{aligned}
&(\lambda y + (1-\lambda)y')\, J_{\mathcal{F}}\left(K\frac{\lambda x + (1-\lambda)x'}{\lambda y + (1-\lambda)y'}\right) \\
=&(\lambda y + (1-\lambda)y')\, J_{\mathcal{F}}\left(\frac{\lambda y}{\lambda y + (1-\lambda)y'}\frac{Kx}{y} + \frac{(1-\lambda)y'}{\lambda y + (1-\lambda)y'}\frac{Kx'}{y'}\right) \\
\geq&\lambda y J_{\mathcal{F}}\left(K\frac{x}{y}\right) + (1-\lambda)y' J_{\mathcal{F}}\left(K\frac{x'}{y'}\right).
\end{aligned}$$

We can therefore claim that $f(x, y) = yJ_{\mathcal{F}}(K\frac{x}{y})$ is concave on $[0, \infty) \times (0, \infty)$. Moreover $f(x, y)$ is non-decreasing in x as $J_{\mathcal{F}}$ is non-decreasing. We also have $f(x, y) = y\int_0^{K\frac{x}{y}} \zeta(\varepsilon)d\varepsilon = x\int_0^1 \zeta\left(K\frac{x}{y}\varepsilon\right)d\varepsilon$.

Since ζ in nonincreasing, f is non-decreasing in y. Finally, because $u \mapsto \sqrt{u}$ is concave, we have

$$\begin{aligned}
&\sqrt{\lambda y + (1-\lambda)y'} J_{\mathcal{F}}\left(K \frac{\sqrt{\lambda x + (1-\lambda)x'}}{\sqrt{\lambda y + (1-\lambda)y'}}\right) \\
=& f\left(\sqrt{\lambda x + (1-\lambda)x'}, \sqrt{\lambda y + (1-\lambda)y'}\right) \\
\geq& f\left(\lambda\sqrt{x} + (1-\lambda)\sqrt{x'}, \lambda\sqrt{y} + (1-\lambda)\sqrt{y'}\right) \\
\geq& \lambda f\left(\sqrt{x}, \sqrt{y}\right) + (1-\lambda) f\left(\sqrt{x'}, \sqrt{y'}\right) \\
=& \lambda\sqrt{y} J_{\mathcal{F}}\left(K \frac{\sqrt{x}}{\sqrt{y}}\right) + (1-\lambda)\sqrt{y'} J_{\mathcal{F}}\left(K \frac{\sqrt{x'}}{\sqrt{y'}}\right).
\end{aligned}$$

4.7.3.3.2 Proof of Lemma S4.11

1. A ball of radius ε for the norm $\lambda||.||$ is a ball of radius ε/λ for the norm $||.||$.

2. A minimal ε-covering for $||.||'$ is also an ε-covering for $||.||$.

3. Consider a minimal $\varepsilon/2$-covering of $\mathcal{H}'$. This is not an $\varepsilon/2$-covering of $\mathcal{H}$ in general because the centers of the covering balls need not be in $\mathcal{H}$. However, in each ball that intersects $\mathcal{H}$, we can select an element of $\mathcal{H}$ as a center of a new ball of radius ε. We thus obtain a new family of balls which forms an ε-covering of $\mathcal{H}$.

4. Let $f_1, ..., f_{N(\varepsilon/2, \mathcal{F}, ||.||)}$ the centers of balls of a minimal $\varepsilon/2$-covering of $\mathcal{F}$. Consider balls of center $f_i - f_j$ and of radius ε for $1 \leq i, j \leq N(\varepsilon/2, \mathcal{F}, ||.||)$. The latter constitute an ε-covering of $\mathcal{F}_\infty$ because for $(g_1, g_2) \in \mathcal{F} \times \mathcal{F}$ we have

$$||(f_i - f_j) - (g_1 - g_2)|| \leq ||f_i - g_1|| + ||f_j - g_2||,$$

which is smaller than ε for at least one pair (i, j).

5. Let $f_1, ..., f_{N(\varepsilon||F||_{Q,2}, \mathcal{F}, ||.||)}$ the centers of balls of a minimal $\varepsilon||F||_{Q,2}$-covering of $\mathcal{F}$ for $||.||_{Q,2}$. Consider balls of center $f_i - f_j$ and radius $8\varepsilon||F^2||_{Q,1}$ for the norm $||.||_{Q,1}$. For every pair $(g_1, g_2) \in \mathcal{F} \times \mathcal{F}$, the Cauchy-Schwarz inequality implies

$$\begin{aligned}
||(f_i - f_j)^2 - (g_1 - g_2)^2||_{Q,1} &\leq ||f_i - f_j + g_1 - g_2||_{Q,2} \times ||(f_i - f_j) - (g_1 - g_2)||_{Q,2} \\
&\leq 4||F||_{Q,2} \times \left(||f_i - g_1||_{Q,2} + ||f_j - g_2||_{Q,2}\right),
\end{aligned}$$

which is smaller than $8\varepsilon||F||_{Q,2}^2 = 8\varepsilon||F^2||_{Q,1}$ for at least one pair (i, j).

Chapter 5

On the construction of confidence intervals for ratios of expectations[1]

Abstract

In econometrics, many parameters of interest can be written as ratios of expectations. The main approach to construct confidence intervals for such parameters is the delta method. However, this asymptotic procedure yields intervals that may not be relevant for small sample sizes or, more generally, in a sequence-of-model framework that allows the expectation in the denominator to decrease to 0 with the sample size. In this setting, we prove a generalization of the delta method for ratios of expectations and the consistency of the nonparametric percentile bootstrap. We also investigate finite-sample inference and show a partial impossibility result: nonasymptotic uniform confidence intervals can be built for ratios of expectations but not at every level. Based on this, we propose an easy-to-compute index to appraise the reliability of the intervals based on the delta method. Simulations and an application illustrate our results and the practical usefulness of our rule of thumb.

Keywords: delta method, confidence regions, uniformly valid inference, sequence of models, nonparametric percentile bootstrap.

Based on [58] : Derumigny, A., Girard, L., & Guyonvarch Y., On the construction of confidence intervals for ratios of expectations. *Arxiv preprint*, arXiv:1904.07111, 2019.

5.1 Introduction

In applied econometrics, the prevalent method for constructing confidence intervals (CIs) is asymptotic: the theoretical guarantees for most CIs used in practice hold only when the number of observations tends to infinity. For a large class of parameters, the construction of asymptotic CIs also relies on the delta method. In this paper, we focus on parameters that can be expressed as ratios of expectations for which the delta method is a standard procedure to conduct inference. The objective is twofold: study the behavior of the delta method and other confidence intervals in some difficult settings and provide tools to detect cases in which the delta method may behave poorly.

[1] Note to the referees: This chapter is based on an arxiv working paper that is still preliminary. It has only been presented internally at CREST.

Many popular parameters in economics take the form of ratios of expectations. Typical examples are conditional expectations since any conditional expectation with a discrete conditioning variable, or a conditioning event, can be written as a ratio of unconditional expectations. For instance, assume that we observe an independent and identically distributed (i.i.d.) sample of individuals indexed by $i \in \{1, \dots, n\}$ with W_i the wage of an individual and D_i an indicator equal to 1 whenever individual i belongs to some treatment group, say a training program; 0 otherwise. Suppose you are interested in the average wage of participants in the program. We have $\mathbb{E}[W \mid D = 1] = \mathbb{E}[WD] / \mathbb{E}[D]$ as D is binary.

Most confidence intervals used in practice are based on asymptotic justifications, hence possible concerns as regards their finite-sample reliability. For ratios of expectations, we document this issue on simulations (see Section 5.3.1). One of our findings is that the coverage of the CIs based on the delta method happens to be far below their nominal level, even for large sample sizes, when the expectation in the denominator is close to 0.[2] For some scenarios, these asymptotic CIs require above 100,000 observations to get reasonably close to their nominal level. Yet, denominators close to 0 are not unusual in practice. Coming back to the treatment/wage example, a small denominator would correspond to a binary treatment with a low participation rate.

In order to deal with that issue, we consider sequences of models, namely we authorize the distribution of the observations to change with the sample size. This framework enables to formalize in an asymptotic way the idea of a denominator close to 0. Indeed, in a standard asymptotic viewpoint, with the expectation in the denominator different from 0, all parameters are fixed and well-defined. Hence, n always grows large enough so that empirical means are close to their expectations and the CIs based on the delta method are valid. In other words, the signal that we want to estimate is constant while the noise goes to 0, and therefore the problem vanishes in this asymptotic perspective. We would like to model more difficult cases, in which the signal can go to 0 as well. This is precisely what the sequence-of-model set-up allows.[3] This is similar to some frameworks that have been developed for weak instrumental variables (IV), see notably [129, 130, 8].

In this literature, another approach does not consider sequences of models but designs "robust" procedures that allow to be exactly in the problematic case, namely a null covariance between the instrument and the endogenous regressor (see [6]). In this case, the parameter of interest is unidentified. In contrast with the weak IV framework, it is worth noting that for ratios in general the parameter of interest is not even defined when the denominator is exactly equal to 0. As a consequence, such an approach seems difficult to extend to our problem.

In our setting, it is unclear, even asymptotically, what the properties of the CIs based on the delta method are when the expectation in the denominator tends to 0. We show that usual CIs can fail and the limiting law of $\widehat{\theta}_n - \theta_n$ may not be Gaussian anymore, denoting by θ_n the ratio of expectations and $\widehat{\theta}_n$ its empirical counterpart. In some cases, the difference $\widehat{\theta}_n - \theta_n$ may actually have a Cauchy limit, as can be found in the weak IV literature.

We show in this sequence-of-model framework that confidence intervals provided by the nonparametric percentile bootstrap have the same asymptotic properties as the ones obtained with the delta method. Simulations support that claim and even suggest the former have better coverage than the latter in finite samples.

Even in standard settings with a fixed but small denominator, simulations document that asymptotic-

[2] The definitions of coverage and other fundamental properties of confidence intervals are recalled in Section 5.8 with the conventions that we use.

[3] This can also rationalize the practice of applied social researchers (see Example 5.1). The heuristic idea is that researchers can consider narrower effects as the data gets richer.

based CIs may require very large sample sizes to attain their nominal level. This suggests to study more in details nonasymptotic inference. More precisely, we construct finite-sample CIs, extending old-established concentration inequalities for means to ratios of means. Concentration inequalities for the mean refer to upper bounds on the probability that an empirical mean departs from its expectation more than a given threshold. Such inequalities permit to construct confidence intervals valid for any sample size and for large classes of probability distributions (see in particular [25]). To our knowledge, there is no such result for ratios. We consider distributions within a class characterized by a lower bound on the first moment for the denominator variable, and an upper bound on the second moment for both the numerator and denominator variables.[4]

One additional result highlights there exists a critical confidence level, above which it is not possible to construct nonasymptotic CIs, uniformly valid on such classes, and that are almost surely of finite length under every distribution of those classes. More precisely, we exhibit explicit upper and lower bounds on this critical confidence level: the former is a threshold above which we show it is impossible to construct such CIs; the latter is a threshold below which we show how to construct them.

These ideas closely relate to some impossibility results as regards the construction of confidence intervals. A large share of the research effort has concentrated on the problem of constructing confidence intervals for expectations. In an early contribution, [14] show that, when $\mathcal{P}$ is the set of all distributions on the real line with finite expectation, the parameter of interest $\theta(P)$ is the expectation with respect to a distribution $P \in \mathcal{P}$ and $\Theta = \mathbb{R}$, a confidence interval built from an i.i.d. sample of $n \in \mathbb{N}^*$ observations that has uniform coverage $1-\alpha$ over $\mathcal{P}$ must contain any real number with probability at least $1-\alpha$. Broadly speaking, any confidence interval must have infinite length with positive probability for every $P \in \mathcal{P}$ to ensure a coverage of $1-\alpha$.

Stronger results can be derived when one further restricts $\mathcal{P}$ or Θ. When $\mathcal{P}$ is taken to be the set of all distributions on the real line with variance uniformly bounded by a finite constant, it is possible to show (using the Bienaymé-Chebyshev inequality) that for every $n \in \mathbb{N}^*$ and every $\alpha \in (0,1)$, there exists a confidence interval that is almost surely of bounded length under every $P \in \mathcal{P}$ and has coverage $1-\alpha$. In this case, the obtained CIs have the advantage that their length shrinks to 0 at the optimal rate $1/\sqrt{n}$. But on the downside, they are not of size $1-\alpha$, even asymptotically, except for some extreme distributions. This means that they tend to be conservative in practice.

A strand of the literature has also investigated more complex problems in which $\theta(P)$ is not restricted to being an expectation. For general parameters, [63] derives a generalization of [14]. An implication of the results in [63] is the existence of an impossibility theorem for ratios of expectations. Let P be a distribution on $\mathbb{R}^2$ with marginals P_X and P_Y. If $\theta(P) = \mathbb{E}_{P_X}[X] / \mathbb{E}_{P_Y}[Y]$, then for every $\alpha \in (0,1)$, it is impossible to build nontrivial CIs of coverage $1-\alpha$ when $\mathcal{P}$ is the set of all distributions on $\mathbb{R}^2$ with finite second moments and $\Theta = \{\theta = \mathbb{E}_{P_X}[X] / \mathbb{E}_{P_Y}[Y] : (\mathbb{E}_{P_X}[X], \mathbb{E}_{P_Y}[Y]) \in \mathbb{R} \times \mathbb{R}^*\}$. As will be explained below, this impossibility result disappears as soon as $\mathcal{P}$ is chosen such that $|\mathbb{E}_{P_Y}[Y]|$ is bounded away from 0 uniformly over $\mathcal{P}$. Interestingly, the impossibility breaks down only partly in the sense that there remains an upper bound on confidence levels (that depends on n) above which it is impossible to build nontrivial CIs.

Other interesting results can be found in [123] and [120]. [123] construct nonasymptotic valid confidence intervals that happen to be also asymptotically optimal. However, they only consider expectations. [120] study smooth functions of a vector of means and give bounds on the distance between the distribution of the normalized and centered estimator and its Gaussian limiting distribution. Nonetheless, the authors do

[4] We refer to this setting as the "Bienaymé-Chebyshev" (BC) case. In Section 5.10, we present similar results for distributions whose supports are bounded ("Hoeffding" case).

not link their results to the construction of confidence intervals.

In the light of that existing literature, our nonasymptotic findings can be interpreted as a partial impossibility result. Indeed, even if we assume a known positive lower bound on the expectation in the denominator, the limitation on the attainable coverage of our nonasymptotic CIs remains. That point complements [63]: for a given sample size n, interesting CIs can be built but not at every confidence level. By contrast, provided the expectation in the denominator is not null, the delta method gives CIs at every confidence level, but their coverage is only asymptotic.

To bridge this gap, we suggest a rule of thumb to assess the reliability of the delta method for ratios of expectations in finite samples. The heuristic idea is simply, for a given sample, to compute an estimator of the lower bound on the above-mentioned critical confidence level. This lower bound can be seen as a conservative value for the unknown critical level, which is a necessary criterion to conduct valid inference in finite samples uniformly over a given class of distributions. Hence, for any desired level higher than this bound, the CIs based on the delta method cannot reach this desired uniform level in finite samples. We illustrate the empirical usefulness of that rule of thumb on simulations and with an application to gender wage disparities in France for the years 2010-2017.

The rest of the paper is organized as follows. Section 5.2 details our framework and assumptions. In Section 5.3, we illustrate the weaknesses of the CIs based on the delta method with a denominator "close to 0" on simulations and detail the asymptotic behavior of the delta method and of the nonparametric percentile bootstrap in our sequence-of-model setting. Section 5.4 is devoted to the construction of nonasymptotic confidence intervals and presents a lower bound on the aforementioned critical confidence level. In Section 5.5, we derive an upper bound on the critical confidence level as well as a lower bound on the length of nonasymptotic CIs. This section also includes the description of a practical index to gauge the soundness of the CIs based on the delta method in finite samples. Section 5.6 present simulations and an application to a real dataset to illustrate our methods. Section 5.7 concludes. General definitions about confidence intervals are recalled in Section 5.8. The proofs of all results are postponed to Section 5.9. Additional results under an alternative set of assumptions ("Hoeffding" case) are detailed in Section 5.10. Section 5.11 presents supplementary simulations.

5.2 Our framework

Throughout the paper, for any random variable U and n i.i.d. replications $(U_{1,n}, \ldots, U_{n,n})$, we denote by $\overline{U}_n$ the empirical mean of U, that is $n^{-1}\sum_{i=1}^{n} U_{i,n}$. Assumption 5.1 defines our sequence-of-model framework and provides the basic requirements to state our asymptotic results.

Assumption 5.1. *For every $n \in \mathbb{N}^*$, we observe a sample $(X_{i,n}, Y_{i,n})_{i=1,\ldots,n} \overset{\text{i.i.d.}}{\sim} P_{X,Y,n}$, where $P_{X,Y,n}$ is a given distribution on $\mathbb{R}^2$ that satisfies $\mathbb{E}[Y_{1,n}] > 0$, $\mathbb{E}\left[X_n^2\right] < +\infty$, and $\mathbb{E}\left[Y_n^2\right] < +\infty$.*

Remark that n indexes both the distribution $P_{X,Y,n}$ of the observations in this model and the number of observations n. This encompasses the standard i.i.d. set-up if the distribution does not change with n: for every $n \in \mathbb{N}^*$, $P_{X,Y,n} = P_{X,Y}$ for some given distribution $P_{X,Y}$. As we assume the existence of a finite expectation, we can consider $\mathbb{E}[Y_{1,n}] \geq 0$ without loss of generality.[5] In order to have properly defined ratios of interest, we need to assume away a null denominator, namely suppose that for every $n \in \mathbb{N}^*$, $\mathbb{E}[Y_{1,n}] > 0$.

Example 5.1 (Sequences of models and the practice of applied researchers)**.**
Researcher may look at the average value of a variable $A_{i,n}$ of interest in a subgroup of the data.

[5] Otherwise, we simply replace $Y_{i,n}$ by its opposite $-Y_{i,n}$.

Subgroups could be defined as the intersections of, say, time, geographical area, gender, age, income brackets and so on. As the number of observations n grows, it is possible to consider subgroups g_n that become thinner and thinner (intersection of more and more variables for instance). This practice could be modelled as estimating $\theta_n := \mathbb{E}[A_{i,n} \mid G_{i,n} = 1] = \mathbb{E}[A_{i,n}G_{i,n}] / \mathbb{P}(G_{i,n} = 1)$ where $G_{i,n}$ is a binary variable that is equal to 1 if an individual i belongs to the subgroup g_n. This corresponds to our framework denoting $X_{i,n} := A_{i,n} \times G_{i,n}$ and $Y_{i,n} := G_{i,n}$.

To derive our nonasymptotic results, Assumption 5.1 has to be strengthened.

Assumption 5.2. *For every $n \in \mathbb{N}^*$, there exist positive finite constants $l_{Y,n}$, $u_{X,n}$, and $u_{Y,n}$ such that (i) $\mathbb{E}[Y_{1,n}] \geq l_{Y,n} > 0$, (ii) $\mathbb{E}\left[X_n^2\right] \leq u_{X,n}$ and $\mathbb{E}\left[Y_n^2\right] \leq u_{Y,n}$.*

Note that in practice, the value of the constants $l_{Y,n}$, $u_{X,n}$, and $u_{Y,n}$ may not be available for practitioners. This is the reason why, in Section 5.5.3, we propose heuristic methods that palliate the lack of knowledge of those constants.

The first part of the assumption bounds the expectation of $Y_{1,n}$ away from 0 while the second states that the second moments of $X_{1,n}$ and $Y_{1,n}$ are bounded. These are necessary to derive nonasymptotic CIs with maintained coverage uniformly over a class of distributions and that are not trivial. Otherwise, if $l_{Y,n} = 0$ or in the absence of the upper bounds $u_{X,n}$ and $u_{Y,n}$, the impossibility theorem of [63] applies and prevents from constructing nontrivial CIs for any confidence level. In a way, given this result, Assumption 5.2 can be seen as close to the minimal hypothesis that allows for the possibility of nontrivial confidence intervals with finite-sample guarantees for ratios of expectations. Furthermore, the sequence-of-model framework allows $l_{Y,n}$ to decrease to 0, which enables us to study limiting cases close to but different from the problematic case $l_{Y,n} = 0$.

This set-up, where Assumptions 5.1 and 5.2 hold, is named the *BC case* since it is possible under these assumptions to construct nonasymptotic CIs using the Bienaymé-Chebyshev inequality. In Section 5.10, we present an adapted version of our results under the assumption that $X_{1,n}$ and $Y_{1,n}$ have a bounded support instead of bounded second moments; a setting we call the *Hoeffding case*.

To sum up, Assumptions 5.1 and 5.2 define a set $\mathcal{P}$ of distributions for some constants $l_{Y,n}$, $u_{X,n}$ and $u_{Y,n}$. For a distribution $P_{X,Y,n}$ in $\mathcal{P}$, the parameter of interest $\theta(P_{X,Y,n})$ is denoted $\theta_n := \mathbb{E}[X_{1,n}] / \mathbb{E}[Y_{1,n}]$ with values in $\mathbb{R}$. To estimate this parameter, we consider its empirical counterpart $\widehat{\theta}_n := \overline{X}_n / \overline{Y}_n$. We seek to construct confidence intervals $C_{n,\alpha}$ for θ_n with nominal level $1 - \alpha$ based on this estimator.

In practice, it is possible that $\overline{Y}_n = 0$ and it may even happen with a strictly positive probability for non-continuous distributions of Y. The estimator $\widehat{\theta}_n$ is not well-defined for such samples. How can we construct a confidence interval $C_{n,\alpha}$ using $\widehat{\theta}_n$ in that context? We could choose to define $C_{n,\alpha} = \mathbb{R}$. This entails that θ_n belongs to $C_{n,\alpha}$ by construction. We believe that such a choice would artificially improve the coverage of $C_{n,\alpha}$ as it induces that the higher $\mathbb{P}(\overline{Y}_n = 0)$, the better the interval in terms of coverage. As a result, we adopt the convention that $\widehat{\theta}_n = +\infty$ and $C_{n,\alpha} = \emptyset$ meaning that we reject the hypothesis $\theta_n = \theta_0$ for every $\theta_0 \in \mathbb{R}$ using the duality between tests and confidence intervals.

5.3 Limitations of the delta method: when are asymptotic confidence intervals valid?

In practice, for a sample of size n, the coverage of asymptotic CIs may be well below their nominal level $1 - \alpha$. Intuitively, this phenomenon should be driven by "problematic" distributions in $\mathcal{P}$ in the

following sense: when the true distribution P is close to the boundary of the class $\mathcal{P}$, the probability $c(n, P) := \mathbb{P}_{P^{\otimes n}}\left(C_{n,\alpha} \ni \theta(P)\right)$ may be much smaller than $1-\alpha$.[6]

In Section 5.3.1, with $C_{n,\alpha}$ the confidence interval based on the delta method, we illustrate on simulations that $c(n, P)$ can fail to match $1-\alpha$ when the expectation in the denominator is fixed close to 0. In other words, it may require a very large number of observations to make reasonable the asymptotic approximation. In Section 5.3.2, we investigate a more serious issue: in the sequence-of-model framework, we let the expectation in the denominator not only be small but converge to 0 as n increases. We show on simulations that depending on the speed at which the denominator goes to 0, $c(n, P)$ can either converge to the nominal level (more or less quickly) or even not converge at all to this target. This sheds light on a partial failure of the delta method when the denominator goes to 0 that we derive formally in Section 5.3.3. Finally, in Section 5.3.4, we show the asymptotic consistency of the nonparametric percentile bootstrap (also known as Efron's percentile bootstrap) in this sequence-of-model framework.

5.3.1 Asymptotic approximation takes time to hold

In this subsection, we consider the i.i.d. case.[7] Under Assumption 5.1, asymptotic confidence intervals are easily obtained combining the multivariate central limit theorem (CLT) and the delta method:

$$\sqrt{n}\left(\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}[X]}{\mathbb{E}[Y]}\right) \xrightarrow[n\to+\infty]{d} \mathcal{N}\left(0, \Sigma\right), \tag{5.1}$$

where $\Sigma = \mathbb{V}[X]/\,\mathbb{E}[Y]^2 + \mathbb{E}[X]^2\mathbb{V}[Y]/\,\mathbb{E}[Y]^4 - 2\,\mathbb{C}ov\left[X, Y\right]\mathbb{E}[X]/\,\mathbb{E}[Y]^3$ and in practice is replaced by a consistent estimate (Slutsky's lemma).

To assess the quality of the CI based on (5.1), we compute its $c(n, P)$ using simulations for different sample sizes n and distributions P and compare it to the nominal level. By definition, the pointwise coverage $c(n, P)$ forms an upper bound on the uniform coverage. In our simulations, we choose the level $1-\alpha = 95\%$. For different sample sizes n and values of $\mathbb{E}[Y]$, we draw $M =$ 5,000 i.i.d. samples of size n following $\mathcal{N}(1,1) \otimes \mathcal{N}(\mathbb{E}[Y], 1)$. We compute $c(n, P)$ for the interval based on the delta method for every pair $(n,\ \mathbb{E}[Y])$ using the 5,000 replications. The expectation $\mathbb{E}[Y]$ ranges from 0.01 (the denominator is close to 0) to 0.75 (the denominator is far from 0). Figure 5.1 sums up the results. For every n, it turns out that the closer $\mathbb{E}[Y]$ to 0, the smaller the $c(n, P)$ of the delta method. When $\mathbb{E}[Y] = 0.01$, we observe that $c(n, P)$ gets close to the nominal level only for n above 300,000. Additional simulations indicate that the phenomenon is robust across different choices of the distribution $P_{X,Y}$ (see Section 5.11).

5.3.2 Asymptotic results may not hold in the sequence-of-model framework

Unlike the result displayed in (5.1), it is unclear how $\sqrt{n}\left(\overline{X}_n/\,\overline{Y}_n - \mathbb{E}[X]/\,\mathbb{E}[Y]\right)$ behaves asymptotically when we consider sequences of models such that the expectation in the denominator tends to 0 as n increases. For a given specification, Figure 5.2 shows the $c(n, P)$ of the CIs based on the delta method when $\mathbb{E}\left[Y_{1,n}\right] = Cn^{-b}$ where C is set to 0.025 and b varies. For a speed $b \geq 1/2$ (i.e. faster than the usual rate of the CLT), the pointwise coverage $c(n, P)$ of the asymptotic CIs obtained by (5.1) is not good in the sense that it is far lower than the nominal level $1-\alpha$ and it does not converge to the latter. Our simulations even suggest that the coverage tends to 0 for $b > 1/2$. For $b < 1/2$, the upper bound $c(n, P)$

[6] Recall that in the nonasymptotic approach, the *coverage* of any given confidence interval $C_{n,\alpha}$ is defined as the infimum of $c(n, P)$ for P ranging over the studied class $\mathcal{P}$ of distributions.

[7] For every $n \in \mathbb{N}^*$, $P_{X,Y,n}$ is identical, hence denoted $P_{X,Y}$. To simplify notations, we also denote by (X, Y) a random vector following $P_{X,Y}$.

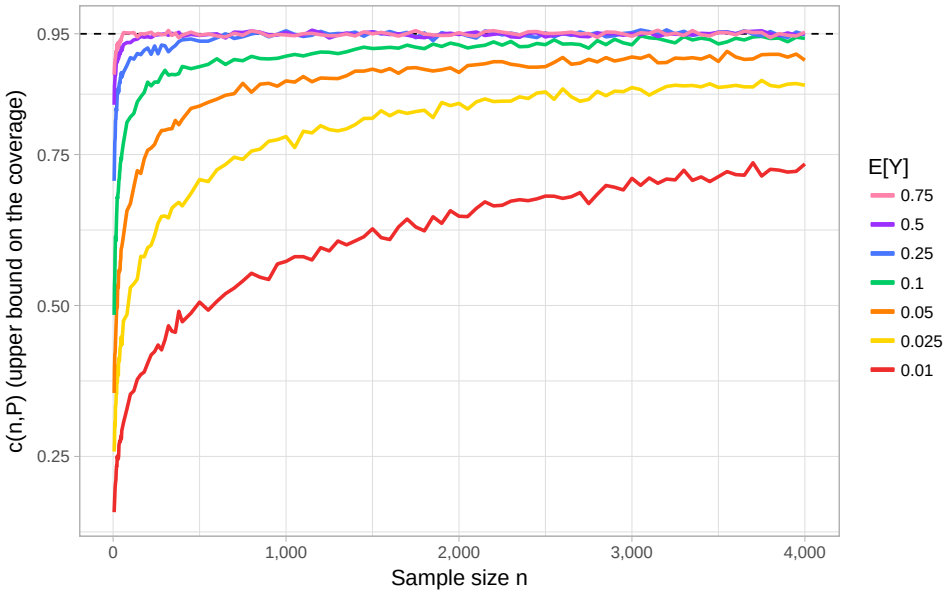

Figure 5.1 – $c(n, P)$ of the asymptotic CIs based on the delta method as a function of the sample size n. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}(1,1) \otimes \mathcal{N}(\mathbb{E}[Y], 1)$. The nominal pointwise asymptotic level is set to 0.95. For each pair $(\mathbb{E}[Y], n)$, the coverage is obtained as the mean over 5,000 repetitions.

on the coverage of the delta method seems to tend to $1 - \alpha$. Yet, in line with Figure 5.1, the validity of the asymptotic approximation requires very large sample sizes.

At this stage, Figure 5.2 presents some evidence that the CIs based on the delta method need to be adapted for sequences of models and that the rate of decrease toward 0 of the expectation $\mathbb{E}[Y_{1,n}]$ matters. The next subsection details formal results in this set-up.

5.3.3 Extension of the delta method for ratios of expectations in the sequence-of-model framework

We are interested in the asymptotic distribution, as n tends to infinity, of the real random variable $S_n := \sqrt{n}\left(\overline{X}_n / \overline{Y}_n - \mathbb{E}[X_{1,n}] / \mathbb{E}[Y_{1,n}]\right)$. The following theorem states the asymptotic behavior of S_n according to the comparison of $\mathbb{V}[Y_{1,n}] / \sqrt{n}$ and $\mathbb{E}[Y_{1,n}]$ under a multivariate Lyapunov condition. It is proved in Section 5.9.1.

We show that in some cases $|S_n| \xrightarrow{a.s.} +\infty$. It is then impossible to state the limiting distribution S_n in the traditional sense. Despite that, we can still get a more precise result looking at the subsequent terms in the asymptotic expansion of S_n. Such an asymptotic expansion is complicated to state, especially in our sequence-of-model framework, since the distributions $P_{X,Y,n}$ change with n without any link from one to the next. To overcome this problem, we consider equivalents in distribution of S_n in the following sense. We say that two sequences of random variables S_n and T_n are *equivalent in distribution* if there exist a probability space $\tilde{\Omega}$ and two sequences of random variables $\tilde{S}_n, \tilde{T}_n$ such that $\forall n \in \mathbb{N}^*$, $S_n \overset{d}{=} \tilde{S}_n$ and $T_n \overset{d}{=} \tilde{T}_n$, and $\tilde{S}_n$ is equivalent to $\tilde{T}_n$ almost surely as $n \to \infty$. This means that for almost every $\tilde{\omega} \in \tilde{\Omega}$, $\tilde{S}_n(\tilde{\omega})$ is equivalent to $\tilde{T}_n(\tilde{\omega})$ (considered as deterministic sequences of real numbers). This notion enables to formalize the link between S_n and a simpler expression T_n.

Theorem 5.1. *Let Assumption 5.1 hold and (i)* $\mathbb{V}[(\gamma_{X,n} X_{1,n}\ ,\ \gamma_{Y,n} Y_{1,n})] \to V$ *as* $n \to \infty$ *for some positive sequences* $\{\gamma_{X,n}\}_{n \in \mathbb{N}^*}$ *and* $\{\gamma_{Y,n}\}_{n \in \mathbb{N}^*}$ *where* V *is a definite positive* 2×2 *matrix,*
(ii) $\sup_{n \in \mathbb{N}^*} \mathbb{E}\left[|X_{1,n}|^3 \gamma_{X,n}^3 + |Y_{1,n}|^3 \gamma_{Y,n}^3\right] < +\infty$*, and (iii)* $\mathbb{P}(\overline{Y}_n = 0) \to 0$ *as* $n \to \infty$*.*

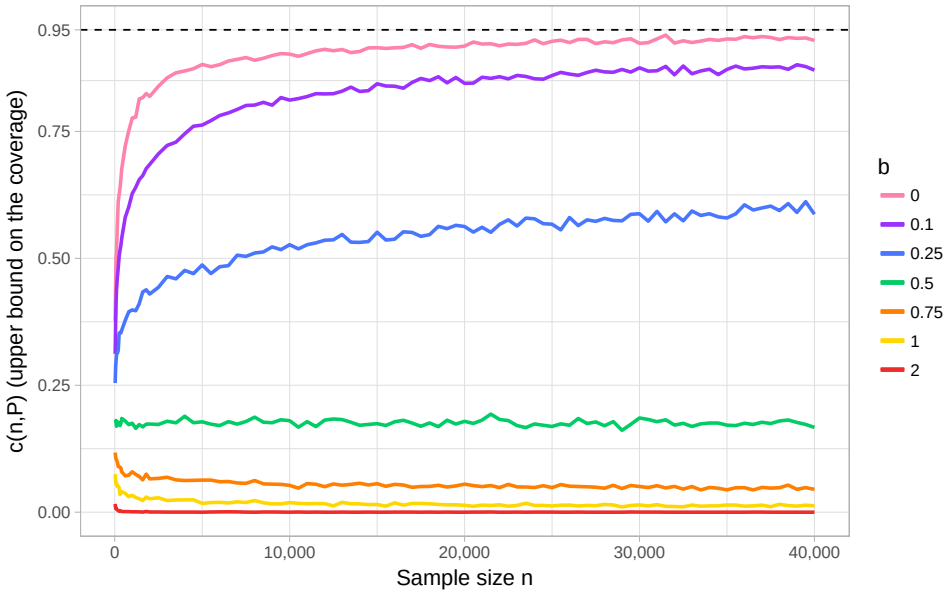

Figure 5.2 – $c(n, P)$ of the asymptotic CIs based on the delta method as a function of the sample size n. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}(1,1) \otimes \mathcal{N}(Cn^{-b}, 1)$, with $C = 0.025$. The nominal pointwise asymptotic level is set to 0.95. For each pair (b, n), the coverage is obtained as the mean over 5,000 repetitions.

Denote the signal-to-noise-ratio by $\text{SNR}_n := \mathbb{E}[Y_{1,n}] / (V_{2,2}^{1/2} n^{-1/2} \gamma_{Y,n}^{-1})$.

Then, the sequence of random variables $S_n := \sqrt{n}\left(\overline{X}_n / \overline{Y}_n - \mathbb{E}[X_{1,n}] / \mathbb{E}[Y_{1,n}]\right)$ *satisfies as* $n \to \infty$*:*

1. *If* $\text{SNR}_n \to +\infty$*, then* S_n *is equivalent in distribution to:*

$$\frac{\sqrt{n}\gamma_{X,n}(\overline{X}_n - \mathbb{E}[X_{1,n}])}{\mathbb{E}[Y_{1,n}]\,\gamma_{X,n}} - \frac{\sqrt{n}\gamma_{Y,n}(\overline{Y}_n - \mathbb{E}[Y_{1,n}])\mathbb{E}[X_{1,n}]}{\mathbb{E}[Y_{1,n}]^2\,\gamma_{Y,n}}.$$

2. *If there exists a finite constant* $C \neq 0$ *such that* $\text{SNR}_n \to C$*, then* S_n *is equivalent in distribution to:*

$$n\gamma_{Y,n}\mathbb{E}[X_{1,n}]\left(\frac{1}{C + \sqrt{n}\gamma_{Y,n}(\overline{Y}_n - \mathbb{E}[Y_{1,n}])} - \frac{1}{C}\right) + \frac{n\gamma_{X,n}(\overline{X}_n - \mathbb{E}[X_{1,n}]) \times \gamma_{Y,n}}{\left(C + \sqrt{n}\gamma_{Y,n}(\overline{Y}_n - \mathbb{E}[Y_{1,n}])\right) \times \gamma_{X,n}}.$$

3. *If* $\text{SNR}_n \to 0$*, then* S_n *is equivalent in distribution to:*

$$\sqrt{n}\left(\frac{\sqrt{n}\gamma_{X,n}(\overline{X}_n - \mathbb{E}[X_{1,n}])}{\sqrt{n}\gamma_{Y,n}(\overline{Y}_n - \mathbb{E}[Y_{1,n}])} \times \frac{\gamma_{Y,n}}{\gamma_{X,n}} - \frac{\mathbb{E}[X_{1,n}]}{\mathbb{E}[Y_{1,n}]}\right).$$

Theorem 5.1 can thus be interpreted as a generalization of the result given by the CLT and the delta method for ratios of expectations. The sequence-of-model framework allows both the expectation and the variance in the denominator to tend to 0. In particular, this happens whenever $Y_{i,n}$ follows a Bernoulli distribution with a parameter p_n tending to 0, as detailed in Example 5.2. For instance, when we estimate a conditional expectation with a discrete conditioning variable or a conditioning event, the denominator is an average of indicator variables that follow a Bernoulli distribution. Figure 5.3 and its companion table highlight the different asymptotic regimes depending on the behaviors of $\{\mathbb{E}[X_{1,n}]\}_{n\in\mathbb{N}^*}$, $\{\mathbb{E}[Y_{1,n}]\}_{n\in\mathbb{N}^*}$, $\{\gamma_{X,n}\}_{n\in\mathbb{N}^*}$ and $\{\gamma_{Y,n}\}_{n\in\mathbb{N}^*}$.

The main takeaway of Theorem 5.1 is that when $\mathbb{E}[X_{1,n}] = C_1/n^a$, $\mathbb{E}[Y_{1,n}] = C_2/n^b$ and $\mathbb{V}[Y] = C_3/n^{b'}$ for some constants $C_1, C_2, C_3 \neq 0$, and $b < 1/2 + b'$, S_n properly renormalized by n to some power still

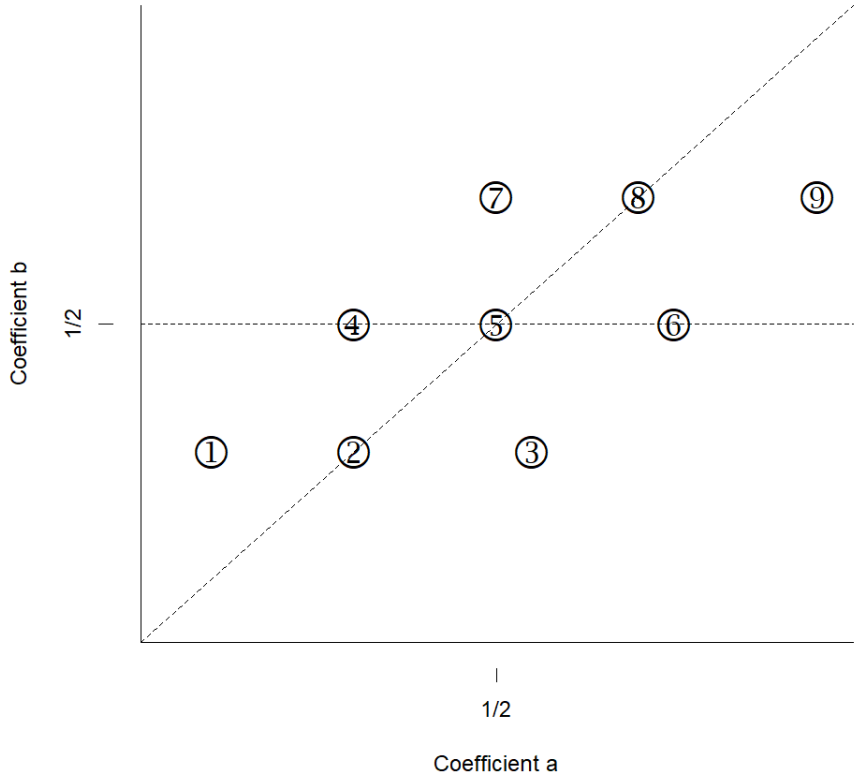

Figure 5.3 – Separation between the different asymptotic regimes as a function of (a, b) for fixed $(a', b') = (0, 0)$, in the case where $\mathbb{E}[X_{1,n}] = C_1/n^a$, $\mathbb{V}[X] = 1/n^{a'}$, $\mathbb{E}[Y_{1,n}] = C_2/n^b$, and $\mathbb{V}[Y] = 1/n^{b'}$, $(a, a', b, b') \in \mathbb{R}_+^4$.

	$a + b' < b + a'$	$a + b' = b + a'$	$a + b' > b + a'$
$b > 1/2 + b'$	$n^{1/2+b'-a'} W_1/W_2$	$n^{1/2+b'-a'}\left(W_1/W_2 - C_1/C_2\right)$	$-n^{1/2+b-a} C_1/C_2$
$b = 1/2 + b'$	$n^{1-a+b'}\left(C_1/(C_2 + W_2) - C_1/C_2\right)$	$n^{1/2+b'-a'}\left(C_1/(C_2 + W_2)\right.$ $\left. - C_1/C_2 + W_1/(C_2 + W_2 n^{a'})\right)$	$n^{1/2+b'-a'}\left(W_1/(C_2 + W_2 n^{a'})\right)$
$b < 1/2 + b'$	$n^{2b-a-b'} C_1 W_2/C_2^2$	$n^{b-a'}(W_1/C_1 - C_1 W_2/C_2^2)$	$n^{b-a'} W_1/C_1$

Table 5.1 – Limiting law of $S_n := \sqrt{n}\left(\overline{X}_n/\overline{Y}_n - \mathbb{E}[X_{1,n}]/\mathbb{E}[Y_{1,n}]\right)$ in the nine different regimes. The couple of variables (W_1, W_2) follow the distribution $\mathcal{N}(0, V)$, where $V = \lim_{n\to+\infty} \mathbb{V}\left[(n^{a'} X_{1,n}, n^{b'} Y_{1,n})\right]$.

converges in distribution to a Normal random variable. This can be explained using the signal-to-noise ratio (SNR) defined in Theorem 5.1. Indeed, in this first case, the SNR_n tends to $+\infty$: the signal in the denominator (that is the expectation of $Y_{1,n}$) is asymptotically bigger than the noise (which is $1/(\gamma_{Y,n}n^{1/2})$ up to a constant factor). Asymptotic inference based on the Normal approximation remains valid, even if the length of such confidence intervals may not decrease with the sample size n.

In all other cases, when the noise dominates in the denominator, S_n converges weakly to a non-Gaussian distribution, in some cases to a generalized Cauchy distribution with parameters that depend on the data generating process (up to a normalization of some power of n). By construction, when the noise dominates, we do not have much information and thus may not be able to conduct inference in these settings. This echoes the impossibility results presented in Section 5.5. In the next section, we provide another method for constructing confidence intervals using the nonparametric percentile bootstrap.

Example 5.2. *When $Y_{1,n}$ follows a Bernoulli distribution with parameter p_n in $(0,1)$, we are always in the first case of Theorem 5.1, meaning that its expectation p_n is always larger than the noise $\sqrt{p_n(1-p_n)/n}$. This latter formula is obtained by remarking that the standard deviation of $Y_{i,n}$ is $\sqrt{p_n(1-p_n)}$ so that $\gamma_{Y,n} = 1/\sqrt{p_n(1-p_n)}$. However, in order to satisfy the constraint $\mathbb{P}(\overline{Y}_n = 0) \to 0$, we have to impose that $np_n \to +\infty$. Therefore, when $p_n = n^{-b}$, confidence intervals based on the delta method will be pointwise consistent if $b < 1$.*

5.3.4 Validity of the nonparametric bootstrap for sequences of models

In this part, we construct confidence intervals for ratios of expectations using Efron's percentile bootstrap. This technique relies on the nonparametric bootstrap resampling scheme that we now recall. We fix a number $B > 0$ of bootstrap replications. For a given initial sample $(X_{i,n}, Y_{i,n}), i = 1, \dots, n$, and a given integer b smaller than B, we define the bootstrapped sample $(X_{i,n}^{(b)}, Y_{i,n}^{(b)}), i = 1, \dots, n$, which is obtained by n i.i.d. resampling from the initial sample, i.e. with replacement. Let $\overline{X}_n^{(b)} := n^{-1}\sum_{i=1}^n X_{i,n}^{(b)}$ be the empirical mean of the numerator in the b-th bootstrapped sample (resp. $\overline{Y}_n^{(b)}$ for the denominator).

Then, Efron's percentile bootstrap, also known as the nonparametric percentile bootstrap, consists in using the quantiles of the bootstrapped distribution conditional on the data to conduct inference. More precisely, for every $\tau \in (0,1)$, let q_τ^{boot} denote the quantile at level τ of $\overline{X}_n^{(1)}/\overline{Y}_n^{(1)}$, which is estimated in practice by the empirical quantile at level τ of the bootstrapped statistics $\left(\overline{X}_n^{(b)}/\overline{Y}_n^{(b)}\right)_{b=1,\dots,B}$. For a given nominal level $1-\alpha \in (0,1)$, the confidence interval we consider is defined as $C_{n,\alpha}^{boot} := \left[q_{\alpha/2}^{boot}\,,\, q_{1-\alpha/2}^{boot}\right]$. The following theorem states the consistency of this interval. It is proved in Section 5.9.2.

Theorem 5.2. *Let Assumption 5.1 hold and (i) $\mathbb{V}[(\gamma_{X,n}X_{1,n}\,,\,\gamma_{Y,n}Y_{1,n})] \to V$ as $n \to \infty$ for some positive sequences $\{\gamma_{X,n}\}_{n\in\mathbb{N}^*}$ and $\{\gamma_{Y,n}\}_{n\in\mathbb{N}^*}$ where V is a definite positive 2×2 matrix,*
(ii) $\sup_{n\in\mathbb{N}^*}\mathbb{E}\Big[(\gamma_{X,n}X_{1,n})^{4+\delta} + (\gamma_{Y,n}Y_{1,n})^{4+\delta}\Big] < +\infty$ for some $\delta > 0$, (iii) $\mathbb{P}(\overline{Y}_n = 0) \to 0$ as $n \to \infty$, and (iv) $\mathbb{P}(\overline{Y}_n^{(1)} = 0) \to 0$ as $n \to \infty$.
Denote the signal-to-noise-ratio by $\text{SNR}_n := \mathbb{E}[Y_{1,n}]/(V_{2,2}^{1/2}n^{-1/2}\gamma_{Y,n}^{-1})$.

If $\text{SNR}_n \to +\infty$, then for every $\alpha \in (0,1)$, the confidence interval $C_{n,\alpha}^{boot}$ is pointwise consistent at level $1-\alpha$, viz. $\mathbb{P}\big(C_{n,\alpha}^{boot} \ni \mathbb{E}[X_{1,n}]/\mathbb{E}[Y_{1,n}]\big) \to 1-\alpha$ as $n \to \infty$.

The assumption $\mathbb{P}(\overline{Y}_n^{(1)} = 0) \to 0$ is satisfied for a large set of cases, for instance when the variables $Y_{i,n}$ are continuous or when they follow a Bernoulli distribution with a parameter decreasing to 0 not too fast (see Example 5.3 below).

Note that the moment condition of order $4+\delta$ is nearly sharp. Indeed, the proofs require the strong law of large numbers for $n^{-1}\sum_{i=1}^n X_{1,n}^2$ and $n^{-1}\sum_{i=1}^n Y_{1,n}^2$. As we are dealing with a triangular array of

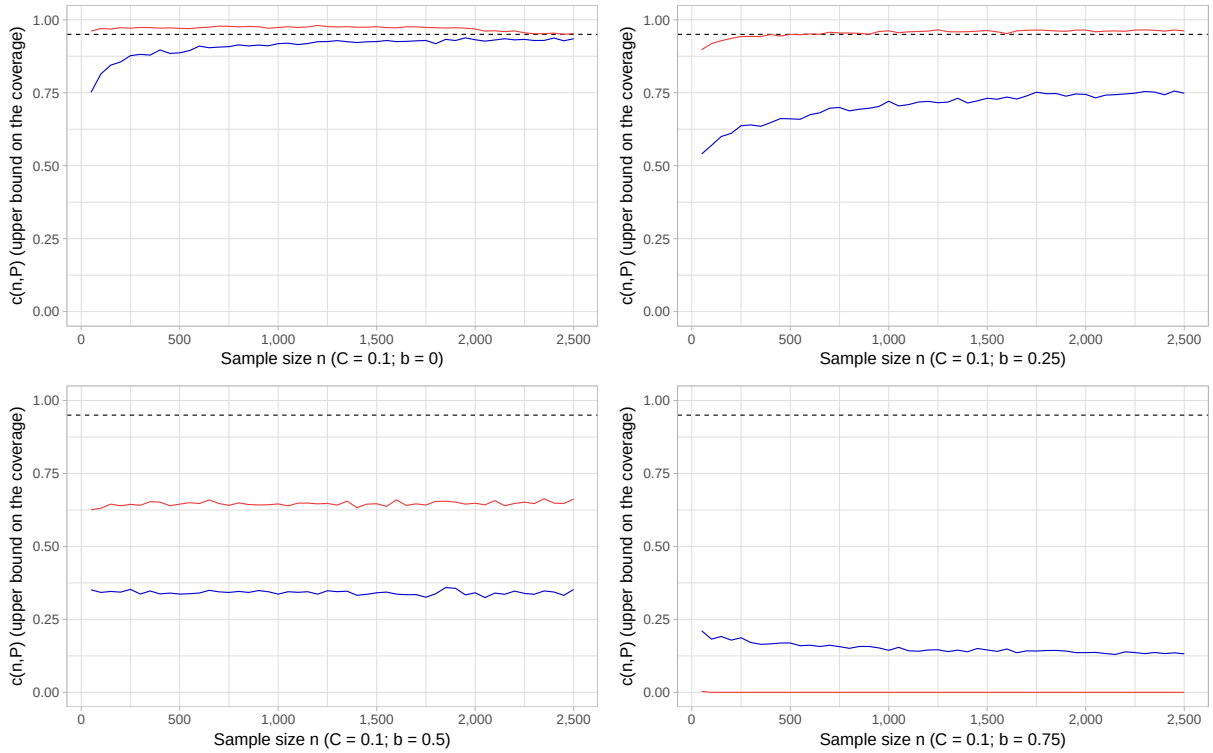

Figure 5.4 – $c(n, P)$ of the asymptotic CIs based on the delta method (blue) and of the CIs constructed with Efron's percentile bootstrap using 2,000 bootstrap replications (red).
Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}(1,1) \otimes \mathcal{N}(Cn^{-b}, 1)$, with $C = 0.1$ and $b \in \{0, 0.25, 0.5, 0.75\}$. The nominal pointwise asymptotic level is set to 0.95. For each pair (b, n), the coverage is obtained as the mean over 5,000 repetitions.

random variables, Theorem 3.1 of [84] shows that moments of order at least 4 are necessary, even in the simpler case where the distribution $P_{X,Y,n}$ does not depend on n.

Example 5.3 (Example 5.2 continued)**.** *When $Y_{1,n}$ follows a Bernoulli distribution with parameter $p_n = 1/n^b$ for a given $b > 0$, the condition $\mathbb{P}(\overline{Y}_n^{(1)} = 0) \to 0$ is satisfied when $b < 1$. We refer the reader to Section 5.9.3 for a proof of this claim.*

In practice, even if the theoretical results of the delta method and of the bootstrap are valid under nearly the same set of assumptions, we observe in the simulations in Figure 5.4 a gap between their pointwise coverage.[8] This fact appears even when $P_{X,Y,n}$ does not depend on n (i.e. $b = 0$). Nonetheless, the coverage gap between these two methods shrinks as n increases provided $b < 0.5$. In the sequence of models where the denominator decreases slowly (i.e. $b = 0.25$) in Figure 5.4, the bootstrap's coverage is much higher than the one of the delta method. Therefore, the CI provided by the nonparametric percentile bootstrap may be an interesting alternative compared to the delta method when conducting inference with a given sample. This is all the more so as the mean in the denominator is close to 0 (in Figure 5.4, of the size of $n^{-0.25}/10$ for a variance normalized to 1) and the number of observations is moderately large (a few thousands here).

[8] Additional simulations comparing the two types of asymptotic confidence intervals are presented in Section 5.11.7.

5.4 Construction of nonasymptotic confidence intervals for ratios of expectations

To construct nonasymptotic confidence intervals, we rely on the possibility to ensure that with large probability (i) $\overline{X}_n$ is close to $\mathbb{E}[X_{1,n}]$, and (ii) $\overline{Y}_n$ is both close to $\mathbb{E}[Y_{1,n}]$ and bounded away from 0. Under Assumptions 5.1 and 5.2, the Bienaymé-Chebyshev inequality can be applied to obtain (i) and (ii). On the other hand, without further restrictions, we are only able to build nonasymptotic CIs at nominal levels that are not too close to 1 (see Section 5.4.2).

This limitation does not arise with nonasymptotic confidence intervals for expectations. In that sense, we can say that building nonasymptotic CIs for ratios of expectations is more demanding. Intuitively, the extra difficulty of the latter task comes from the need to ensure (ii). To stress that point, we show in the next subsection that when $\overline{Y}_n$ is bounded away from 0 and positive almost surely, we can build nonasymptotic CIs at every nominal level.

5.4.1 An easy case: the support of the denominator is well-separated from 0

We present a simple framework in which it is possible to build nonasymptotic CIs, valid for every $n \in \mathbb{N}^*$, and with coverage $1-\alpha$ for every $\alpha \in (0,1)$. To do so, we restrict further the set $\mathcal{P}$ of admissible distributions with the following assumption.

Assumption 5.3. *For every $n \in \mathbb{N}^*$, there exists a positive finite constant $a_{Y,n}$ such that $Y_{1,n} \geq a_{Y,n}$ almost surely.*

Under Assumption 5.3, for every $n \in \mathbb{N}^*$, $\overline{Y}_n \geq a_{Y,n} > 0$ almost surely under every distribution in $\mathcal{P}$ and $\overline{Y}_n^{-1}$ is bounded from above. This assumption obviously rules out binary $\{0,1\}$ random variables in the denominator of the ratio, which can be quite restrictive in practice. Under this assumption, the following theorem gives a concentration inequality for our ratio of expectations. It is proved in Section 5.9.4.

Theorem 5.3. *Let Assumptions 5.1, 5.2 and 5.3 hold. For every $n \in \mathbb{N}^*$, $\varepsilon > 0$, we have*

$$\sup_{P \in \mathcal{P}} \mathbb{P}_{P^{\otimes n}}\left(\left| \frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}[X_{1,n}]}{\mathbb{E}[Y_{1,n}]} \right| > \frac{\left(\varepsilon + \sqrt{u_{X,n}}\right)\varepsilon}{a_{Y,n} l_{Y,n}} + \frac{\varepsilon}{l_{Y,n}} \right) \leq \frac{u_{X,n}}{n\varepsilon^2} + \frac{u_{Y,n} - l_{Y,n}^2}{n\varepsilon^2}.$$

As a consequence, $\inf_{P \in \mathcal{P}} \mathbb{P}_{P^{\otimes n}}\left(\mathbb{E}[X_{1,n}] / \mathbb{E}[Y_{1,n}] \in \left[\overline{X}_n / \overline{Y}_n \pm t \right] \right) \geq 1-\alpha$*, with the choice*

$$t := \frac{1}{l_{Y,n}} \sqrt{\frac{u_{X,n} + u_{Y,n} - l_{Y,n}^2}{n\alpha}} \left(1 + \frac{1}{a_{Y,n}} \left\{ \sqrt{\frac{u_{X,n} + u_{Y,n} - l_{Y,n}^2}{n\alpha}} + \sqrt{u_{X,n}} \right\} \right),$$

for every $\alpha \in (0,1)$.

The theorem shows that it is possible to construct nonasymptotic CIs for ratios of expectations, with guaranteed coverage at every confidence level, that are almost surely of bounded length under every distribution in $\mathcal{P}$ characterized by Assumptions 5.1, 5.2 and 5.3. In Section 5.4.2, we give an analogous result that only requires Assumptions 5.1 and 5.2 to hold, so that it encompasses the case of $\{0,1\}$-valued denominators. However, the cost to pay will be an upper bound on the achievable coverage of the confidence intervals.

5.4.2 General case: no assumption on the support of the denominator

We seek to build nontrivial nonasymptotic CIs under Assumptions 5.1 and 5.2 only. Under Assumption 5.1, $\mathbb{E}[Y_{1,n}] \neq 0$, so that there is no issue in considering the fraction $\mathbb{E}[X_{1,n}]/\mathbb{E}[Y_{1,n}]$. However, without Assumption 5.3, $\{\overline{Y}_n = 0\}$ has positive probability in general so that $|\overline{X}_n/\overline{Y}_n| < +\infty$ with probability less than one. Note that when $P_{Y,n}$ is continuous with respect to Lebesgue's measure, $\overline{X}_n/\overline{Y}_n$ is finite with probability one anymore since the event $\{\overline{Y}_n = 0\}$ has probability zero. This is not an easier case from a theoretical point of view though since, without more restrictions, $\overline{Y}_n$ can still be arbitrarily close to 0 with positive probability.

Theorem 5.4. *Let Assumptions 5.1 and 5.2 hold. For every $n \in \mathbb{N}^*$, $\varepsilon > 0, \tilde{\varepsilon} \in (0,1)$, we have*

$$\sup_{P\in\mathcal{P}} \mathbb{P}_{P^{\otimes n}}\left(\left|\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}[X_{1,n}]}{\mathbb{E}[Y_{1,n}]}\right| > \left(\frac{\left(\sqrt{u_{X,n}}+\varepsilon\right)\tilde{\varepsilon}}{(1-\tilde{\varepsilon})^2} + \varepsilon\right)\frac{1}{l_{Y,n}}\right) \leq \frac{u_{X,n}}{n\varepsilon^2} + \frac{u_{Y,n}-l_{Y,n}^2}{n\tilde{\varepsilon}^2 l_{Y,n}^2}.$$

As a consequence, $\inf_{P\in\mathcal{P}} \mathbb{P}_{P^{\otimes n}}\left(\mathbb{E}[X_{1,n}]/\mathbb{E}[Y_{1,n}] \in \left[\overline{X}_n/\overline{Y}_n \pm t\right]\right) \geq 1-\alpha$*, with the choice*

$$t = \frac{1}{l_{Y,n}}\left(\frac{\left(\sqrt{u_{X,n}} + \sqrt{2u_{X,n}/(n\alpha)}\right)\sqrt{2(u_{Y,n}-l_{Y,n}^2)/(n\alpha l_{Y,n}^2)}}{\left(1-\sqrt{2(u_{Y,n}-l_{Y,n}^2)/(n\alpha l_{Y,n}^2)}\right)^2} + \sqrt{\frac{2u_{X,n}}{n\alpha}}\right),$$

for every $\alpha > \overline{\alpha}_n := \frac{2(u_{Y,n}-l_{Y,n}^2)}{nl_{Y,n}^2}$.[9]

This theorem is proved in Section 5.9.5. It states that when $l_{Y,n} > 0$, it is possible to build valid nonasymptotic CIs with finite length up to the confidence level $1-\overline{\alpha}_n$. This is a more positive result than [63] which states that it is not possible to build nontrivial nonasymptotic CIs when $l_{Y,n}$ is taken equal to 0, no matter the confidence level. Note that Theorem 5.4 is not an impossibility theorem since it only claims that considering confidence levels smaller than $1-\overline{\alpha}_n$ is *sufficient* to build nontrivial CIs under Assumptions 5.1 and 5.2. The remaining question is to find out whether it is *necessary* to focus on confidence levels that do not exceed a certain threshold under Assumptions 5.1 and 5.2. We answer this in Section 5.5.1.

Theorem 5.4 has two other interesting consequences: for every confidence level up to $1-\overline{\alpha}_n$, a nonasymptotic interval of the form $\left[\overline{X}_n/\overline{Y}_n \pm \tilde{t}\right]$ with $\tilde{t} > t$ has coverage $1-\alpha$ but is unnecessarily conservative. Moreover, if the data generating process does not depend on n (i.e. in the standard i.i.d. set-up), the length of the confidence interval shrinks at the optimal rate $1/\sqrt{n}$ for every fixed α. Note that the coefficient 2 in the definition of $\overline{\alpha}_n$ defined above can be reduced to any number $w > 1$, at the expense of increasing the length of the confidence interval (this length actually tends to infinity when w tends to 1).

5.5 Nonasymptotic CIs: impossibility results and practical guidelines

In this section, we prove two impossibility results: a maximum confidence level above which it is impossible to build nontrivial nonasymptotic CIs and a necessary lower bound on the length of nonasymptotic CIs.

[9] Equivalently, it means that for a given α, the above choice of t is valid for every integer $n > \overline{n}_\alpha := 2(u_{Y,n}-l_{Y,n}^2)/(\alpha l_{Y,n}^2)$.

5.5.1 An upper bound on testable confidence levels

Theorem 5.5. *Let $\mathcal{P}$ be the class of all distributions satisfying Assumptions 5.1 and 5.2 and $\underline{\alpha}_n := \left(1 - l_{Y,n}^2/u_{Y,n}\right)^n$. Let $n \in \mathbb{N}^*$, and a random set I_n that satisfies $I_n = \emptyset$ whenever $\overline{Y}_n = 0$. Then $\sup_{P\in\mathcal{P}} \mathbb{P}_{P^{\otimes n}}\left(I_n = \emptyset\right) \geq \underline{\alpha}_n$.*

This theorem is proved in Section 5.9.6. Combining the latter result and Theorem 5.4, we conclude that there exists some critical level $1 - \alpha_n^c$ belonging to the interval $[1 - \overline{\alpha}_n, 1 - \underline{\alpha}_n]$ such that it is impossible to build nontrivial nonasymptotic confidence intervals based on $\overline{X}_n/\overline{Y}_n$ if and only if their nominal level is above $1 - \alpha_n^c$. It is worth remarking that with a sample of size n, the CIs based on the delta method with a nominal level $1 - \alpha > 1 - \alpha_n^c$ cannot have coverage $1 - \alpha$ uniformly over $\mathcal{P}$ as such CIs verify the conditions of Theorem 5.5. Finally remark that when $u_{Y,n}/l_{Y,n}^2 = 1$, there is no impossibility result anymore: assume that $u_{Y,n}/l_{Y,n}^2 = 1$ and let Q be a distribution on $\mathbb{R}^2$ that satisfies Assumptions 5.1 and 5.2. Let $(X_{i,n}, Y_{i,n})_{i=1}^n \overset{\text{i.i.d.}}{\sim} Q$. We have that $\mathbb{V}[Y_{1,n}] = 0$, which implies that $Y_{1,n} = \mathbb{E}[Y_{1,n}]$ almost surely. Assumption 5.1 further ensures that $Y_{1,n} \neq 0$ almost surely. Consequently, the results of Section 5.4.1 apply and allow us to conclude that under Assumptions 5.1, 5.2 and $u_{Y,n}/l_{Y,n}^2 = 1$, it is possible to build nontrivial nonasymptotic CIs at every confidence level. Indeed, in that case, we are in fact only estimating a simple mean, and therefore there is no constraint on α.

Figure 5.5 below shows the critical level and its bounds obtained in our nonasymptotic results.

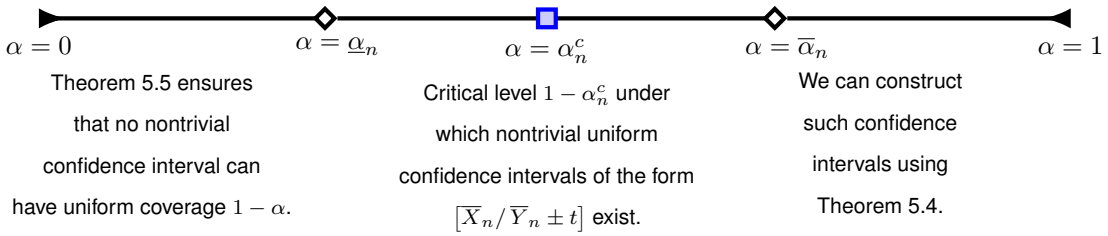

Figure 5.5 – The critical level and its bounds.

In the same spirit as in Theorem 5.1, we could consider a modified version of the signal-to-noise ratio defined by $\widetilde{\text{SNR}}_n := l_{Y,n}/(u_{Y,n}^{1/2} n^{-1/2})$. When we have enough information ($\widetilde{\text{SNR}}_n \to +\infty$), the critical level $1 - \alpha_n^c$ tends to 1. Therefore, for every $\alpha \in (0, 1)$, nonasymptotic confidence intervals can be constructed at every level for n large enough. On the contrary, when $\widetilde{\text{SNR}}_n \to 0$, the critical level $1 - \alpha_n^c$ tends to 0, which means that it is impossible to construct uniformly valid CIs for n large enough. Finally, when $\widetilde{\text{SNR}}_n \to C$ for a positive constant C, a critical level remains as in the nonasymptotic case since $\underline{\alpha}_n \to \exp(-C)$.

5.5.2 A lower bound on the length of nonasymptotic confidence intervals

The following theorem is an extension of [34][Proposition 6.2] to ratios. It is proved in Section 5.9.7.

Theorem 5.6. *For every integer $n \geq 7$, $\alpha \in \left(0, 1 \wedge n/\left(l_{Y,n} + \sqrt{u_{Y,n} - l_{Y,n}^2}\right)^2\right)$, and $\xi < 1$ there exists a distribution Q on $\mathbb{R}^2$ that satisfies Assumptions 5.1 and 5.2 such that for $(X_{i,n}, Y_{i,n})_{i=1}^n \overset{i.i.d}{\sim} Q$, we have*

$$\mathbb{P}_{Q^{\otimes n}}\left(\left|\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}[X_{1,n}]}{\mathbb{E}[Y_{1,n}]}\right| > \xi\sqrt{\frac{v_n}{3n\alpha}}\right) > \alpha,$$

where $v_n := u_{X,n}/\left(l_{Y,n} + \sqrt{u_{Y,n} - l_{Y,n}^2}\right)^2$.

With this theorem, we can claim that CIs of the form $\left[\overline{X}_n/\overline{Y}_n \pm t\right]$ cannot have uniform coverage $1-\alpha$, for every $\alpha \in \left(0, 1 \wedge n/\left(l_{Y,n} + \sqrt{u_{Y,n} - l_{Y,n}^2}\right)^2\right)$, under Assumptions 5.1 and 5.2 if they are shorter than $\sqrt{v_n/(3n\alpha)}$. By a careful inspection of the proof (see Lemma 5.8), we can in fact replace the value 3 in the theorem by any number strictly larger than $e = \exp(1)$, at the price of assuming $n \geq n_0$ for n_0 large enough. It is interesting to note that the distributions Q that are built in the proof of the theorem are on the boundary of $\mathcal{P}$ in the sense that they satisfy $\mathbb{E}\left[X_n^2\right] = u_{X,n}$, $\mathbb{E}\left[Y_{1,n}\right] = l_{Y,n}$ and $\mathbb{E}\left[Y_n^2\right] = u_{Y,n}$.

5.5.3 Practical methods and plug-in estimators

Nonasymptotic confidence intervals and the thresholds $\overline{\alpha}_n$ and $\overline{n}_\alpha$ based on Theorem 5.4 rely on Assumptions 5.1 and 5.2. In practice, building such CIs or computing those thresholds require the knowledge of the constants $l_{Y,n}$, $u_{X,n}$ and $u_{Y,n}$ that determine the class of distributions we consider.[10] Therefore, we need to state some values for those constants. Note that constructing nontrivial and nonasymptotic CIs that overcome the limitations of having to choose some a priori class of distributions is not possible. Indeed, we would get back to [14] and [63] type impossibility results.

How to choose $l_{Y,n}$, $u_{X,n}$ and $u_{Y,n}$ depends on the specific application. Sometimes, stating values can be sensible if researchers do have control or expert knowledge of the variables. Resuming an example started in the introduction, if the variable in the denominator is an indicator of being treated in the setting of a Randomized Controlled Trial, researchers can have intuitions about reasonable values for the lower and upper bounds of the probability of being treated.

The unknown constants are upper and lower bounds on moments that characterize the class $\mathcal{P}$. As such, they can never be recovered from the data since observations are by construction drawn from a single distribution $P \in \mathcal{P}$. Under i.i.d. sampling, sample means converge to their corresponding theoretical moments, provided the latter are finite. Hence, without prior information, a plug-in strategy has to be used which consists in: (i) using the moments of a single distribution instead of the bounds on the class, (ii) estimating those moments with their empirical counterparts. As a consequence, this approach is valid pointwise only and not uniformly over $\mathcal{P}$ anymore. Furthermore, it is only asymptotically justified. On the other hand, for any sample provided $\overline{Y}_n \neq 0$, this plug-in strategy enables us to construct our CIs and the quantity $\overline{n}_\alpha$ (or $\overline{\alpha}_n$), which can be a useful rule of thumb as explained below. We stick to that principle in our simulations and application.

For a given level $1-\alpha$ and a class of distributions satisfying Assumptions 5.1 and 5.2, $\overline{n}_\alpha$ is the minimal sample size required to construct our nonasymptotic CIs. In other words, for a sample size $n < \overline{n}_\alpha$, the data is not rich enough to construct the nonasymptotic CIs of Theorem 5.4 at this level. Heuristically, the comparison of $\overline{n}_\alpha$ and n can be used as a rule of thumb to assess whether the coverage of the CIs based on the delta method matches their nominal level.[11] Several simulations tend to confirm the practical interest of that rule of thumb as $\overline{n}_\alpha$ turns out to be very close to the sample size above which the gap between the coverage of the asymptotic CIs based on the delta method and their nominal level becomes negligible. (see Section 5.6.1 and Section 5.11).

[10] Actually, the computation of $\overline{\alpha}_n$ and $\overline{n}_\alpha$ only require the knowledge of $l_{Y,n}$ and $u_{Y,n}$.

[11] Equivalently, we could compare $\overline{\alpha}_n$ and α. As a rule of thumb, $\overline{\alpha}_n$ can be seen as the lowest α (hence the highest nominal level $1-\alpha$) for which the asymptotic CIs based on the delta method are reliable given the sample size n.

5.6 Numerical applications

5.6.1 Simulations

This section presents simulations that support the use of $\overline{n}_\alpha$, or equivalently $\overline{\alpha}_n$, as a rule of thumb to inspect the reliability of the asymptotic confidence intervals from the delta method.

In Figure 5.6, a nominal level $1-\alpha$ is fixed and we show the $c(n,P)$ of the CIs based on the delta method as a function of the sample size n, as well as $\overline{n}_\alpha$ derived in Theorem 5.4. It happens that the coverage converges toward its nominal level for sample sizes around $\overline{n}_\alpha$, which supports $\overline{n}_\alpha$ as a rule of thumb of interest in practice.[12] In Figure 5.7, a sample size is fixed and we show the coverage for different nominal levels, as well as the quantity $\overline{\alpha}_n$. It is the converse of Figure 5.6 in that sense. In this simulation, $\overline{\alpha}_n$ turns out to fall close to the lowest α (hence highest $1-\alpha$) for which the coverage of the CIs based on the delta method attains their nominal level.

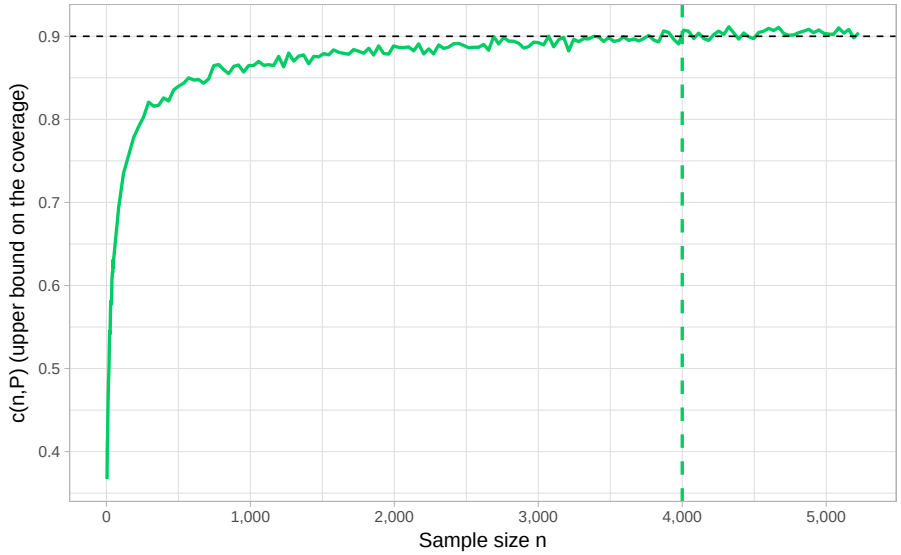

Figure 5.6 – $c(n,P)$ of the asymptotic CIs based on the delta method as a function of the sample size n and $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}_2$ (bivariate Gaussian) with $\mathbb{E}[X] = 0.5$, $\mathbb{E}[Y] = 0.1$, $\mathbb{V}[X] = 1$, $\mathbb{V}[Y] = 2$, $\mathbb{C}orr(X,Y) = 0.5$. The nominal pointwise asymptotic level is set to 0.90. For a sample size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right)/\left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.1$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

All in all, Figures 5.6 and 5.7 and additional simulations advocate the use of $\overline{n}_\alpha$ derived in Theorem 5.4 (or conversely $\overline{\alpha}_n$) as a rule of thumb to appraise the dependability of the CIs obtained with the delta method for ratios of expectations.

5.6.2 Application to real data

We illustrate our methods with an application related to gender wage disparities. The application resumes our canonical example of conditional expectations since we estimate the proportion of women within wage brackets that are defined as having a wage higher than a given threshold. We use $n =$ 204,246 observations from the French Labor Survey data between 2010 and 2017.[13]

[12] This fact holds across various specifications (see additional simulations in Section 5.11).

[13] Enquête Emploi en continu (version FPR) – 2010-2017, INSEE [producteur], ADISP [diffuseur].

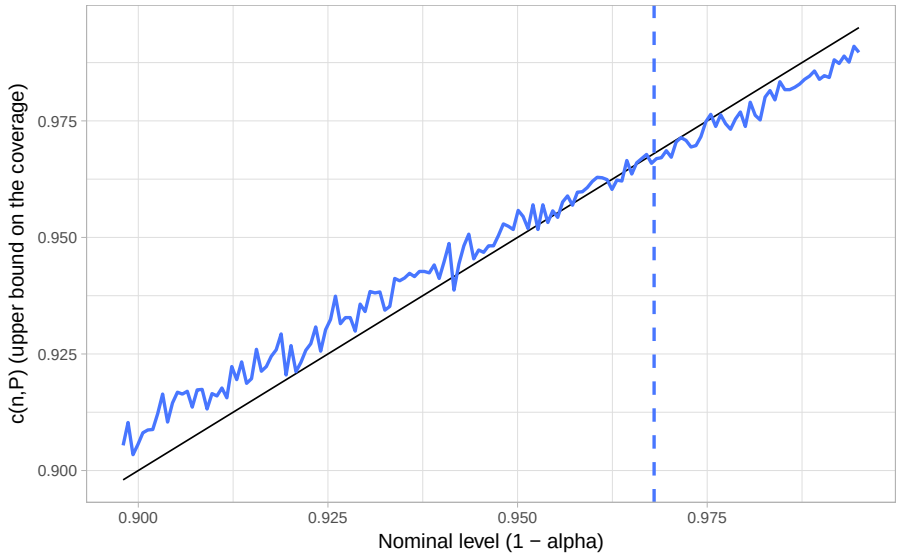

Figure 5.7 – $c(n, P)$ of the asymptotic CIs based on the delta method as a function of the sample size n and $\overline{\alpha}_n$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}_2$ (bivariate Gaussian) with $\mathbb{E}[X] = 0.5$, $\mathbb{E}[Y] = 0.25$, $\mathbb{V}[X] = 2$, $\mathbb{V}[Y] = 1$, $\mathbb{C}orr(X, Y) = 0.5$. The sample size is n = 1,000. For each nominal level $1 - \alpha$ in the x-axis, we draw $10,000$ samples, compute the asymptotic CIs and see whether it covers or not the ratio of interest; we report the mean over the $10,000$ repetitions in the y-axis. The solid line is the first bisector $y = x$. The dashed vertical line shows $\overline{\alpha}_n := 2\left(u_{Y,n} - {l_{Y,n}}^2\right)/\left(n{l_{Y,n}}^2\right)$, setting here $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

Let W be a real random variable that indicates the wage of an employee (expressed in euros per month) and F an indicator variable equal to 1 if the employee is a woman and 0 otherwise. For a given threshold wage w_0, the parameter of interest is $\mathbb{E}[F \mid W \geq w_0]$. It can be written as a ratio of expectations with $X = F\,\mathbb{1}\{W \geq w_0\} = \mathbb{1}\{F = 1, W \geq w_0\}$ in the numerator and $Y = \mathbb{1}\{W \geq w_0\}$ in the denominator. As we consider higher thresholds w_0, the expectation in the denominator gets closer to 0. As an illustration, out of n = 204,246 observations, 355 individuals have monthly wages higher than 10,000 euros (which corresponds to a mean in the denominator equal to 0.0017); 44 individuals above 20,000 ($\overline{Y}_n = 2.2 \times 10^{-4}$); and only 17 above 30,000 ($\overline{Y}_n = 8.3 \times 10^{-5}$).[14]

For various thresholds w_0, Figure 5.8 presents the estimate $\widehat{\theta}_n$ and two 95%-nominal-level confidence intervals for the parameter $\mathbb{E}[F \mid W \geq w_0]$: the one based on the delta method (see Section 5.3.1) and the one using Efron's percentile bootstrap (see Section 5.3.4). With higher thresholds, the expectation in the denominator is closer to 0 which results in wider confidence intervals. For very high thresholds, the CIs become hardly informative. In particular, the lower end of the interval based on the delta method is negative whereas the parameter of interest belongs to $[0, 1]$ by construction.

The dashed vertical line relates to our rule of thumb introduced in Section 5.5.3. More precisely, given the level $1 - \alpha = 0.95$, for each threshold w_0, we compute the plug-in counterpart of $\overline{n}_\alpha$ defined in Theorem 5.4: $2\left(n^{-1}\sum_{i=1}^n Y_i^2 - \overline{Y}_n^2\right)/\left(\alpha\overline{Y}_n^2\right)$. Given that Y is a binary variable, the latter quantity is increasing with w_0 and exceeds n at some threshold represented by the dashed vertical line (here a little above 20,000). Consequently, for higher thresholds, our rule of thumb suggests that the confidence intervals obtained with the delta method might undercover as the expectation in the denominator is "too close to 0" relative to the number of observations. Actually, in the application, it is around this vertical line that the two CIs start to differ. In particular, the upper end of Efron's percentile confidence interval

[14] To give a sense of the wage distribution, note that the empirical quantiles of W at orders 90%; 95%; 99%; and 99.99% are respectively: 2,989; 3,728; 6,000; and 26,024.

Figure 5.8 – Point estimate and confidence intervals for the parameter $\mathbb{E}[F \mid W \geq w_0]$ as a function of the wage threshold w_0. The parameter is the proportion of women within the wage bracket $[w_0, +\infty)$. The nominal level of the CIs is set to 95%. Efron's percentile bootstrap CIs are obtained using 2,000 bootstrap replications. The dashed vertical line represents the lowest wage threshold such that the plug-in counterpart of $\overline{n}_\alpha$ exceeds n.

becomes larger than the upper end of the interval based on the delta method.

5.7 Conclusion

This paper studies the construction of confidence intervals for ratios of expectations, which are frequent parameters of interest in applied econometrics.

The most common method to do so is asymptotic and yields CIs based on the asymptotic normality of the empirical means that estimate the numerator and the denominator combined with the delta method. We document on simulations that the coverage of the confidence intervals based on the delta method may fall short of their nominal level when the expectation in the denominator is close to 0, even with fairly large sample size.

To further study the reliability of those CIs, we use a sequence-of-model framework, analogous to what a strand of the weak IV literature does. Indeed, it enables to consider limiting cases, namely here denominators tending to 0. In the weak IV case, the equivalent is to move closer to a null covariance between the endogenous regressor and the instrument. At the limit, the coefficient of interest is not identified. Our problem differs since the parameter is not even defined in the problematic case of a null denominator. This issue underlies the impossibility type results presented in the paper.

First, in an asymptotic perspective, the possibility of a denominator arbitrarily close to 0 explains why we need a sufficiently slow rate of convergence of the expectation in the denominator to 0 to conduct meaningful inference. More precisely, our main asymptotic results basically show that the CIs based on the delta method are valid, as well as those obtained by Efron's percentile bootstrap, when this speed is lower than $1/\sqrt{n}$ (the standard speed of the CLT). Furthermore, on simulations, Efron's percentile bootstrap CIs reach their nominal level sooner (namely for smaller sample sizes) than the CIs based

on the delta method. It suggests that beyond the sequence-of-model rationalization, when confronted in practice to a mean in the denominator close to 0 relative to the size of the sample at hand, Efron's percentile bootstrap CIs may be more trustworthy than the delta method's ones.

Obviously, those cases where the coverage of the CIs based on the delta method can be well below their nominal level do not self-signal to practitioners. This is why the second part of the paper proposes a rule of thumb to detect those cases and thus assess the dependability of the asymptotic CIs based on the delta method on finite samples. This index is based on the construction of nonasymptotic confidence intervals and on impossibility results that stem from the problematic null denominator case.

In substance, even if we bound away from 0 the expectation in the denominator, there remains a partial impossibility result. Indeed, we show that there exists a critical nominal level above which the coverage of any nonasymptotic confidence interval that is undefined when $\overline{Y}_n = 0$ cannot uniformly attain its target level. More precisely, we derive explicit upper and lower bounds on this critical level as a function of the characteristics of the considered class of distributions. Then, the heuristic of our rule of thumb consists in estimating by plug-in a lower bound on this critical level (or equivalently, for a given level, an upper bound on the minimal required sample size). The resulting index can thus be computed immediately on any sample. In addition to its theoretical foundations, various simulations and an application to real data attest the practical usefulness of this rule of thumb.

This paper can be seen as a first step towards nonasymptotic inference in econometric models where the issue of close-to-zero denominators arises. Notable examples may include weak IV, Wald ratios, and difference-in-difference estimands.

5.8 General definitions about confidence intervals

A standard situation in statistics or econometrics can be modelled as the observation of a sample of $n \in \mathbb{N}^*$ i.i.d. observations valued in some measurable space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$. The statistical model is therefore $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mathcal{P})^{\otimes n}$ with $\mathcal{P}$ some specified set of distributions on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$. For every distribution $P \in \mathcal{P}$, let $\theta(P)$ be a parameter of interest and the map $\theta : P \mapsto \theta(P)$ be valued in a metric space (Θ, d).

We denote by C_n a confidence set for $\theta(P)$. Formally, a confidence set C_n can be defined as a measurable map from $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))^{\otimes n}$ to the measurable space $(\mathcal{F}_\Theta \sqcup \{\text{undefined}\}, \mathcal{B}(\mathcal{F}_\Theta) \sqcup \{\text{undefined}\})$, where $\mathcal{F}_\Theta$ is the family of all closed subsets of Θ and $\mathcal{B}(\mathcal{F}_\Theta)$ is the sigma-algebra generated by $\{F \in \mathcal{F}_\Theta : F \cap K \neq \emptyset\}$ for K running through the family of compact subsets of Θ.

As the vocabulary may somewhat fluctuate between authors, we define below classical objects to fix the notations and terminology used in this paper. The goal is to build confidence sets for a targeted *confidence level* $1-\alpha$ (also termed *nominal level* of the confidence set). For $n \in \mathbb{N}^*$, for $\alpha \in (0,1)$, we say that a confidence set C_n or a sequence of sets $(C_n)_{n\in\mathbb{N}^*}$ has:

i. *coverage* $1-\alpha$ over $\mathcal{P}$ if:
$$\inf_{P\in\mathcal{P}} \mathbb{P}_{P^{\otimes n}}(C_n \ni \theta(P)) \geq 1-\alpha$$

ii. *size* $1-\alpha$ over $\mathcal{P}$ if the inequality is an equality:
$$\inf_{P\in\mathcal{P}} \mathbb{P}_{P^{\otimes n}}(C_n \ni \theta(P)) = 1-\alpha.$$

iii. *asymptotic coverage* $1-\alpha$ *pointwise* over $\mathcal{P}$ if:[15]
$$\forall P \in \mathcal{P}, \liminf_{n\to+\infty} \mathbb{P}_{P^{\otimes n}}(C_n \ni \theta(P)) \geq 1-\alpha.$$

iv. *asymptotic coverage* $1-\alpha$ *uniformly* over $\mathcal{P}$ if:[16]
$$\liminf_{n\to+\infty} \inf_{P\in\mathcal{P}} \mathbb{P}_{P^{\otimes n}}(C_n \ni \theta(P)) \geq 1-\alpha.$$

A confidence set with coverage $1-\alpha$ but size different from $1-\alpha$ over $\mathcal{P}$ is said to be *conservative* over $\mathcal{P}$[17]. We further define a *nontrivial confidence set* as a confidence set that is almost surely strictly included in Θ (whenever it is defined) under every distribution in $\mathcal{P}$. For instance, if $\theta(P)$ is the expectation under P, $\Theta = \mathbb{R}$ and $\mathcal{P}$ is the set of all distributions that admit a finite expectation, a nontrivial CI is any CI that is almost surely of finite length under every distribution in $\mathcal{P}$. For ratios of expectations, $\Theta = \mathbb{R}$ too and we will use the term *almost surely of finite length* as a synonym of nontrivial, without stating "under every distribution in $\mathcal{P}$" when there is no ambiguity as regards the class $\mathcal{P}$ considered.

A family of confidence intervals $(C_{n,\alpha})_{n\in\mathbb{N}^*,\,\alpha\in(0,1)}$ is said to be *pointwise* (resp. *uniformly*) *consistent* if for every $\alpha \in (0,1)$, the sequence $(C_{n,\alpha})_{n\in\mathbb{N}^*}$ has pointwise (resp. uniformly) asymptotic coverage at level $1-\alpha$.

5.9 Proofs of the results in Sections 5.3, 5.4 and 5.5

5.9.1 Proof of Theorem 5.1

Let $\theta_{X,n} := \mathbb{E}[X_{1,n}]$, $\theta_{Y,n} := \mathbb{E}[Y_{1,n}]$. Let $h_{X,n} := \sqrt{n}\gamma_{X,n}(\overline{X}_n - \mathbb{E}[X_{1,n}])$ and $h_{Y,n} := \sqrt{n}\gamma_{Y,n}(\overline{Y}_n - \mathbb{E}[Y_{1,n}])$ be the centered and normalized versions of $\overline{X}_n$ and $\overline{Y}_n$. We first rewrite Theorem 5.1 using this notation.

[15] Respectively *pointwise asymptotic size* when the inequality is replaced by an equality.

[16] Respectively *uniform asymptotic size* when the inequality is replaced by an equality.

[17] Similarly, a confidence set is said to be *asymptotically conservative pointwise* over $\mathcal{P}$ (respectively *uniformly* over $\mathcal{P}$) if property iii. (resp. property iv.) holds with a strict inequality.

Theorem 5.7. *Let Assumption 5.1 hold. Assume that* $\mathbb{V}[(\gamma_{X,n}X_{1,n}\,,\,\gamma_{Y,n}Y_{1,n})] \to V$ *for some positive sequences* $\gamma_{X,n}$ *and* $\gamma_{Y,n}$ *where* V *is a definite positive* 2×2 *matrix, that* $\mathbb{P}(\overline{Y}_n = 0) \to 0$*, as* $n \to \infty$ *and that*

Then the sequence of random variables $A_n := \overline{X}_n/\overline{Y}_n - \theta_{X,n}/\theta_{Y,n}$ *satisfies as* $n \to \infty$*:*

1. *If* $n^{-1/2} = o(\gamma_{Y,n}\theta_{Y,n})$*, then* A_n *is equivalent to*

$$n^{-1/2}\left(\frac{h_{X,n}}{\theta_{Y,n}\gamma_{X,n}} - \frac{h_{Y,n}\theta_{X,n}}{\gamma_{Y,n}\theta_{Y,n}^2}\right).$$

2. *If there exists a finite constant* $C \neq 0$ *such that* $\sqrt{n}\gamma_{Y,n}\theta_{Y,n} \to C$ *as* $n \to \infty$*, then* A_n *is equivalent to*

$$\sqrt{n}\gamma_{Y,n}\theta_{X,n}\left(\frac{1}{C+h_{Y,n}} - \frac{1}{C}\right) + \frac{h_{X,n}\gamma_{Y,n}}{(C+h_{Y,n})\gamma_{X,n}}.$$

3. *If* $\gamma_{Y,n}\theta_{Y,n} = o(n^{-1/2})$*, then* A_n *is equivalent to*

$$\frac{h_{X,n}\gamma_{Y,n}}{h_{Y,n}\gamma_{X,n}} - \frac{\theta_{X,n}}{\theta_{Y,n}}.$$

Let us define $W_n := \mathbb{1}\{\theta_{Y,n} + h_{Y,n}/(\sqrt{n}\gamma_{Y,n}) = 0\}$ and remark that $W_n = 1$ whenever $\overline{Y}_n = 0$. By assumption $\mathbb{P}(\overline{Y}_n = 0) \to 0$, therefore $W_n \xrightarrow[n\to+\infty]{d} \delta_0$. Moreover, by Lyapunov's central limit theorem applied to

$$(h_{X,n}, h_{Y,n}) = \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n}(X_{i,n}\gamma_{X,n}, Y_{i,n}\gamma_{Y,n}) - (\mathbb{E}[X]\gamma_{X,n}, \mathbb{E}[Y]\gamma_{Y,n})\Big),$$

using $V \neq \mathbf{0}$ and the boundedness of $\mathbb{E}\big[|X_{1,n}|^3\big]\gamma_{X,n}^3$ and $\mathbb{E}\big[|Y_{1,n}|^3\big]\gamma_{Y,n}^3$, we obtain $(h_{X,n}, h_{Y,n}) \xrightarrow[n\to+\infty]{d} \mathcal{N}(0,V)$. We also obtain $(h_{X,n}, h_{Y,n}, W_n) \xrightarrow[n\to+\infty]{d} \mathcal{N}(0,V)\otimes\delta_0$ by Slutsky's Lemma. We can therefore apply Skorokhods's almost sure representation theorem, see [136, Theorem 2.19]. It means that there exists a probability space $(\tilde{\Omega}, \tilde{\mathcal{U}}, \tilde{\mathbb{P}})$, a sequence of random vectors $(\tilde{h}_{X,n}, \tilde{h}_{Y,n}, \tilde{W}_n)$ such that for every $n \geq 1$, $(\tilde{h}_{X,n}, \tilde{h}_{Y,n}, \tilde{W}_n) \overset{d}{=} (h_{X,n}, h_{Y,n}, W_n)$, and a random vector $(\tilde{h}_{X,\infty}, \tilde{h}_{Y,\infty}, \tilde{W}_\infty)$ following the distribution $\mathcal{N}(0,V)\otimes\delta_0$ such that $(\tilde{h}_{X,n}, \tilde{h}_{Y,n}, \tilde{W}_n) \xrightarrow{a.s.} (\tilde{h}_{X,\infty}, \tilde{h}_{Y,\infty}, \tilde{W}_\infty)$, where the convergence is to be seen as of a sequence of random vectors defined on $(\tilde{\Omega}, \tilde{\mathcal{U}}, \tilde{\mathbb{P}})$. Let us define

$$\begin{aligned}\tilde{A}_n := \frac{\theta_{X,n} + \tilde{h}_{X,n}/(\sqrt{n}\gamma_{X,n})}{\theta_{Y,n} + \tilde{h}_{Y,n}/(\sqrt{n}\gamma_{Y,n})} - \frac{\theta_{X,n}}{\theta_{Y,n}} &\overset{d}{=} \frac{\theta_{X,n} + h_{X,n}/(\sqrt{n}\gamma_{X,n})}{\theta_{Y,n} + h_{Y,n}/(\sqrt{n}\gamma_{Y,n})} - \frac{\theta_{X,n}}{\theta_{Y,n}} \\ &= \frac{\overline{X}_n}{\overline{Y}_n} - \frac{\theta_{X,n}}{\theta_{Y,n}} = A_n.\end{aligned}$$

Moreover, we have $\tilde{W}_n = \mathbb{1}\{\theta_{Y,n} + \tilde{h}_{Y,n}/(\sqrt{n}\gamma_{Y,n}) = 0\}$ and $\tilde{W}_\infty = 0$ almost surely. We can define

$$\tilde{\Omega}^* = \{\tilde{\omega} \in \tilde{\Omega} : \tilde{W}_n(\tilde{\omega}) \to 0 \text{ and } \exists N > 0, \forall n \geq N, \tilde{h}_{Y,n}(\tilde{\omega}) \neq 0\}.$$

By the almost sure convergence of $(\tilde{h}_{Y,n}, \tilde{W}_n)$, we get $\tilde{\mathbb{P}}(\tilde{\Omega}^*) = 1$, and for every $\tilde{\omega} \in \tilde{\Omega}^*$, $\tilde{W}_n(\tilde{\omega}) = 0$ and $\tilde{h}_{Y,n}(\tilde{\omega}) \neq 0$ for every n large enough. This means that for every given $\tilde{\omega} \in \tilde{\Omega}^*$, and for every n large enough, $\tilde{A}_n$ is well-defined. In the rest of the proof, we will fix such a $\tilde{\omega} \in \tilde{\Omega}^*$, so that all random variables may be considered as deterministic. By the almost sure representation theorem, this means that the equivalents and limits that will be obtained will still be valid in law in the original spaces Ω_n.

First case: We have

$$\begin{aligned}\tilde{A}_n &= \frac{\overline{X}_n}{\overline{Y}_n} - \frac{\theta_{X,n}}{\theta_{Y,n}} = \frac{\theta_{X,n} + \tilde{h}_{X,n}/(\sqrt{n}\gamma_{X,n})}{\theta_{Y,n} + \tilde{h}_{Y,n}/(\sqrt{n}\gamma_{Y,n})} - \frac{\theta_{X,n}}{\theta_{Y,n}} \\ &= \frac{\theta_{X,n} + \tilde{h}_{X,n}/(\sqrt{n}\gamma_{X,n})}{\theta_{Y,n}}\left(1 - \frac{\tilde{h}_{Y,n}}{\sqrt{n}\gamma_{Y,n}\theta_{Y,n}} + O\big((\sqrt{n}\gamma_{Y,n}\theta_{Y,n})^{-2}\big)\right) - \frac{\theta_{X,n}}{\theta_{Y,n}} \\ &\sim \frac{-\theta_{X,n}\tilde{h}_{Y,n}}{\sqrt{n}\gamma_{Y,n}\theta_{Y,n}^2} + \frac{\tilde{h}_{X,n}}{\sqrt{n}\gamma_{X,n}\theta_{Y,n}},\end{aligned}$$

as claimed.

Second case: We have

$$\begin{aligned}\tilde{A}_n &\sim \frac{\theta_{X,n} + \tilde{h}_{X,n}/(\sqrt{n}\gamma_{X,n})}{C/(\sqrt{n}\gamma_{Y,n}) + \tilde{h}_{Y,n}/(\sqrt{n}\gamma_{Y,n})} - \frac{\theta_{X,n}}{C/(\sqrt{n}\gamma_{Y,n})} \\ &= \frac{\sqrt{n}\gamma_{Y,n}\theta_{X,n} + \tilde{h}_{X,n}\gamma_{Y,n}/\gamma_{X,n}}{C + \tilde{h}_{Y,n}} - \frac{\sqrt{n}\gamma_{Y,n}\theta_{X,n}}{C}.\end{aligned}$$

We factorize by $\theta_{X,n}$ in the latter expression, which completes the proof.

Third case: We have

$$\begin{aligned}\tilde{A}_n &= \frac{\theta_{X,n} + \tilde{h}_{X,n}/(\sqrt{n}\gamma_{X,n})}{\theta_{Y,n} + \tilde{h}_{Y,n}/(\sqrt{n}\gamma_{Y,n})} - \frac{\theta_{X,n}}{\theta_{Y,n}} = \frac{\theta_{X,n} + \tilde{h}_{X,n}/(\sqrt{n}\gamma_{X,n})}{\big(\tilde{h}_{Y,n} + o(1)\big)/(\sqrt{n}\gamma_{Y,n})} - \frac{\theta_{X,n}}{\theta_{Y,n}} \\ &\sim \frac{\sqrt{n}\theta_{X,n}\gamma_{Y,n}}{\tilde{h}_{Y,n}} + \frac{\tilde{h}_{X,n}\gamma_{Y,n}}{\tilde{h}_{Y,n}\gamma_{X,n}} - \frac{\theta_{X,n}}{\theta_{Y,n}} \\ &\sim \theta_{X,n}\left(\frac{\sqrt{n}\gamma_{X,n}}{\tilde{h}_{Y,n}} - \frac{1}{\theta_{Y,n}}\right) + \frac{\tilde{h}_{X,n}\gamma_{Y,n}}{\tilde{h}_{Y,n}\gamma_{X,n}},\end{aligned}$$

and the result follows from the fact that $\sqrt{n}\gamma_{X,n}/\tilde{h}_{Y,n}$ is negligible compared to $1/\theta_{Y,n}$.

□

5.9.2 Proof of Theorem 5.2

For $b = 1, 2$, let $h_{X,n} := \sqrt{n}\gamma_{X,n}(\overline{X}_n - \theta_{X,n})$ (resp. $h^{Y,n}$), $S_n := (h_{X,n}, h_{Y,n})'$ and $S_n^{(b)} := (h_{X,n}^{(b)}, h_{Y,n}^{(b)})'$, where $h_{X,n}^{(b)} := \sqrt{n}\gamma_{X,n}(\overline{X}_n^{(b)} - \overline{X}_n)$ is the b-th bootstrap replication of $h_{X,n}$ (resp. $h_{Y,n}^{(b)}$).

Lemma 5.4. *We have* $d_{BL}\left(P_{S_n^{(1)} \mid (X_{i,n}, Y_{i,n})_{i=1}^n}, \mathcal{N}(0, V)\right) \xrightarrow{a.s.} 0.$

By the Central Limit Theorem, we have $S_n \xrightarrow[n\to+\infty]{d} S$ with $S \sim \mathcal{N}(0, V)$ and by Lemma 5.4 (proved in Section 5.9.2.1) and the triangle inequality, we get $d_{BL}\left(P_{S_n^{(1)} \mid (X_l, Y_l)_{l=1}^n}, P_{S_n}\right) \xrightarrow{p} 0$. Combining both results, Lemma 2.2 in [28] gives us

$$d_{BL}\left(P_{(S_n,\, S_n^{(1)},\, S_n^{(2)})},\, P_S^{\otimes 3}\right) \to 0.$$

Let us define $W_n := \mathbb{1}\{\theta_{Y,n} + h_{Y,n}/(\sqrt{n}\gamma_{Y,n}) = 0\}$ and remark that $W_n = 1$ whenever $\overline{Y}_n = 0$. By assumption $\mathbb{P}(\overline{Y}_n = 0) \to 0$, therefore we have $W_n \xrightarrow[n\to+\infty]{d} \delta_0$. We define also $W_n^{(b)} := \mathbb{1}\{\overline{Y}_n^{(b)}\} = \mathbb{1}\{\overline{Y}_n + h_{Y,n}^{(b)}/(\sqrt{n}\gamma_{Y,n}) = 0\}$, so that $W_n^{(1)} = 1$ whenever $\overline{Y}_n^{(b)} = 0$. In the same way as previously, $W_n^{(b)} \xrightarrow[n\to+\infty]{d} \delta_0$ holds by assumption. Let $Z_n = (S_n, W_n, S_n^{(1)}, W_n^{(1)}, S_n^{(1)}, W_n^{(2)})$ be a random vector of size 9, and let Z be a random vector of size 9 following $(P_S \otimes \delta_0)^{\otimes 3}$.

By Slutsky's lemma, we have $d_{BL}\left(P_{Z_n}, P_Z\right) \to 0$ with our new notation. Using Skorokhods's almost sure representation theorem [136, Theorem 2.19], there exists a probability space Ω^+, a sequence of

random vectors $Z_n^+ \in \mathbb{R}^9$ and a vector Z^+ defined on Ω^+ such that $Z_n^+ \xrightarrow{a.s.} Z^+$, $Z_n \overset{d}{=} Z_n^+$ and $Z \overset{d}{=} Z^+$. Let us use the notation

$$\begin{aligned} Z_n^+ &= \left(S_n^+, W_n^+, S_n^{(1)+}, W_n^{(1)+}, S_n^{(1)+}, W_n^{(2)+}\right) \\ &= \left(h_{X,n}^+, h_{Y,n}^+, W_n^+, h_{X,n}^{(1)+}, h_{Y,n}^{(1)+}, W_n^{(1)+}, h_{X,n}^{(2)+}, h_{Y,n}^{(2)+}, W_n^{(2)+}\right) \\ \text{and } Z^+ &= \left(Z_1^+, Z_2^+, Z_3^+\right), \end{aligned}$$

where $S_n^+, S_n^{(1)+}, S_n^{(2)+}$ are random vectors of dimension 2 and Z_1^+, Z_2^+, Z_3^+ are random vectors of dimension 3. We define

$$\begin{aligned} A_n &:= \frac{\overline{X}_n}{\overline{Y}_n} - \frac{\theta_{X,n}}{\theta_{Y,n}} = \frac{\theta_{X,n} + h_{X,n}/(\sqrt{n}\gamma_{X,n})}{\theta_{Y,n} + h_{Y,n}/(\sqrt{n}\gamma_{Y,n})} - \frac{\theta_{X,n}}{\theta_{Y,n}} \\ A_n^{(b)} &:= \frac{\overline{X}_n + h_{X,n}^{(b)}/(\sqrt{n}\gamma_{X,n})}{\overline{Y}_n + h_{Y,n}^{(b)}/(\sqrt{n}\gamma_{Y,n})} - \frac{\overline{X}_n}{\overline{Y}_n} \\ &= \frac{\theta_{X,n} + h_{X,n}/(\sqrt{n}\gamma_{X,n}) + h_{X,n}^{(b)}/(\sqrt{n}\gamma_{X,n})}{\theta_{Y,n} + h_{Y,n}/(\sqrt{n}\gamma_{Y,n}) + h_{Y,n}^{(b)}/(\sqrt{n}\gamma_{Y,n})} - \frac{\theta_{X,n} + h_{X,n}/(\sqrt{n}\gamma_{X,n})}{\theta_{Y,n} + h_{Y,n}/(\sqrt{n}\gamma_{Y,n})}, \end{aligned}$$

and respectively their counterparts A_n^+ and $A_n^{(b)+}$ defined on Ω^+. The following lemma, proved in Section 5.9.2.2, ensures the existence of an event of probability 1 on which every quantity is well-defined.

Lemma 5.5. *There exists an event $\tilde{\Omega} \subset \Omega^+$ such that $\mathbb{P}(\tilde{\Omega}) = 1$ and such that for every $\omega \in \tilde{\Omega}$, and for all n large enough, $h_{Y,n}^+(\omega) \neq 0$, $h_{Y,n}^{(1)+}(\omega) \neq 0$, $h_{Y,n}^{(2)+}(\omega) \neq 0$ and $A_n^+(\omega)$, $A_n^{(1)+}(\omega)$ and $A_n^{(2)+}(\omega)$ are well-defined.*

In the next step, we fix $\omega \in \widetilde{\Omega}$ and let $C := \lim_{n\to+\infty} \theta_{X,n}\gamma_{X,n}/\theta_{Y,n}\gamma_{Y,n}$ and

$$\sigma_n := \sqrt{n}\theta_{Y,n}\left(\gamma_{X,n}\mathbb{1}_{\{C\in\mathbb{R}\}} + \gamma_{Y,n}\theta_{Y,n}/\theta_{X,n}\mathbb{1}_{\{|C|=+\infty\}}\right).$$

We restrict ourselves to the case $n^{1/2}\gamma_{Y,n}\theta_{Y,n} \to +\infty$. Theorem 5.1 therefore yields

$$\sigma_n A_n^+(\omega) = \begin{cases} -Ch_{Y,n}^+(\omega) + h_{X,n}^+(\omega) + o(1) \text{ if } C \in \mathbb{R} \\ -h_{Y,n}^+(\omega) + o(1) \text{ else.} \end{cases} \tag{5.2}$$

Furthermore, the same tools as those used in the proof of Theorem 5.1 plus the fact that $\theta_{Y,n} + h_{Y,n}^+(\omega)/(\sqrt{n}\gamma_{Y,n}) \sim \theta_{Y,n}$ imply

$$\begin{aligned} &\sigma_n A_n^{(b)+}(\omega) \\ &\sim \sigma_n \left(\frac{-\left(\theta_{X,n} + h_{X,n}^+(\omega)/(\sqrt{n}\gamma_{X,n})\right)}{\sqrt{n}\gamma_{Y,n}\left(\theta_{Y,n} + h_{Y,n}^+(\omega)/(\sqrt{n}\gamma_{Y,n})\right)^2} h_{Y,n}^{(b)+}(\omega) \right. \\ &\qquad \left. + \frac{1}{\sqrt{n}\gamma_{Y,n}\left(\theta_{Y,n} + h_{Y,n}^+(\omega)/(\sqrt{n}\gamma_{X,n})\right)} h_{X,n}^{(b)+}(\omega) \right) \\ &\sim \sigma_n \left(\frac{-\left(\theta_{X,n} + h_{X,n}^+(\omega)/(\sqrt{n}\gamma_{Y,n})\right)}{\sqrt{n}\gamma_{Y,n}\theta_{Y,n}^2} h_{Y,n}^{(b)+}(\omega) + \frac{1}{\sqrt{n}\gamma_{X,n}\theta_{Y,n}} h_{X,n}^{(b)+}(\omega) \right). \end{aligned}$$

We can also remark that when $\theta_{X,n} + h_{X,n}^+(\omega)/(\sqrt{n}\gamma_{X,n}) \sim \theta_{X,n}$

$$\sigma_n A_n^{(b)+}(\omega) = \begin{cases} -Ch_{Y,n}^{(b)+}(\omega) + h_{X,n}^{(b)+}(\omega) + o(1) \text{ if } C \in \mathbb{R} \\ -h_{Y,n}^{(b)+}(\omega) + o(1) \text{ else.} \end{cases} \tag{5.3}$$

When $\theta_{X,n} + h^+_{X,n}(\omega)/(\sqrt{n}\gamma_{X,n}) = O\big(h^+_{X,n}(\omega)/(\sqrt{n}\gamma_{X,n})\big)$, we have $C = 0$ and we find again that

$$\sigma_n A_n^{(b)+}(\omega) = h_{X,n}^{(b)+}(\omega) + o(1). \tag{5.4}$$

Let $D_n^+ := \left(-Ch^+_{Y,n} + h^+_{X,n}\right)\mathbb{1}_{\{|C|<+\infty\}} - h^+_{Y,n}\mathbb{1}_{\{|C|=+\infty\}}$ (resp. D_n, $D_n^{(b)}$ and $D_n^{(b)+}$), which corresponds to the dominant terms in Equations (5.2), (5.3) and (5.4) above. By construction of Z_n^+ and Z_n, we have $Z_n^+ \xrightarrow{a.s.} Z^+$, so that the continuous mapping theorem ensures that $\left(D_n^+, D_n^{(1)+}, D_n^{(2)+}\right) \xrightarrow{a.s.} (U_1, U_2, U_3)$, where for every $i \in \{1,2,3\}$, we define $U_i^+ := \left(-CZ^+_{i,2} + Z^+_{i,1}\right)\mathbb{1}_{\{C\in\mathbb{R}\}} - Z^+_{i,2}\mathbb{1}_{\{|C|=+\infty\}}$ where $Z^+_{i,1}$ (resp. $Z^+_{i,2}$) is the first (resp. second) component of the vector Z_i^+. Combining the triangle inequality, Equations (5.2), (5.3) and (5.4), we get

$$\left(\sigma_n A_n^+, \sigma_n A_n^{(1)+}, \sigma_n A_n^{(2)+}\right) \xrightarrow{a.s.} \left(U_1^+, U_2^+, U_3^+\right).$$

Using the fact that for all $n \in \mathbb{N}$ $\left(A_n, A_n^{(1)}, A_n^{(2)}\right) \overset{d}{=} \left(A_n^+, A_n^{(1)+}, A_n^{(2)+}\right)$, we obtain

$$\left(\sigma_n A_n, \sigma_n A_n^{(1)}, \sigma_n A_n^{(2)}\right) \xrightarrow[n\to+\infty]{d} (U_1^+, U_2^+, U_3^+).$$

Therefore, $d_{BL}\Big(P_{\left(\sigma_n A_n,\, \sigma_n A_n^{(1)},\, \sigma_n A_n^{(2)}\right)},\, P_{U_1^+}^{\otimes 3}\Big) \to 0$ as $n \to +\infty$ and $\sigma_n A_n \xrightarrow[n\to+\infty]{d} U_1^+$. Applying Lemma 2.2 of [28], we can conclude that

$$d_{BL}\Big(P_{\sigma_n A_n^{(1)} \,|\, (X_{i,n}, Y_{i,n})_{i=1}^n},\, P_{U_1^+}\Big) \xrightarrow{p} 0.$$

The conclusion follows from Lemma 23.3 in [136].

□

5.9.2.1 Proof of Lemma 5.4

Let $t = (t_X, t_Y)' \in \mathbb{R}^2$, and denote $T_{i,n} = t_X\gamma_{X,n}X_{i,n} + t_Y\gamma_{Y,n}Y_{i,n}$ for $i = 1, \ldots, n$ and $T_{i,n}^{(1)}$ its bootstrap counterpart. Let also $V_{T_{1,n}} := t'\mathbb{V}[(\gamma_{X,n}X_{1,n}\,,\,\gamma_{Y,n}Y_{1,n})]t$ and $V_T := t'Vt$. We start by showing that for every $t \in \mathbb{R}^2$, $P_{\sqrt{n}\left(\overline{T}_n^{(1)} - \overline{T}_n\right) \,|\, (X_{i,n}, Y_{i,n})_{i=1}^n}$ converges weakly to $P_T = \mathcal{N}(0, V_T)$ almost surely conditionally on $(X_{i,n}, Y_{i,n})_{i=1}^n$ in the sense of the Lévy criterion for weak convergence, i.e.

$$\left|\mathbb{E}\left[e^{iu\sqrt{n}\left(\overline{T}_n^{(1)} - \overline{T}_n\right)} \mid (X_{i,n}, Y_{i,n})_{i=1}^n\right] - e^{u^2V_T/2}\right| \xrightarrow{a.s.} 0 \quad \forall u \in \mathbb{R}. \tag{5.5}$$

To do so, we have to check the steps of the proof of Theorem 23.4 in [136]. We have

$$\mathbb{E}\left[\overline{T}_n^{(1)} \mid (X_{i,n}, Y_{i,n})_{i=1}^n\right] = \overline{T}_n \quad \text{and}$$
$$\mathbb{E}\left[\left(T_{i,n}^{(1)} - \overline{T}_n\right)^2 \mid (X_{i,n}, Y_{i,n})_{i=1}^n\right] = \frac{1}{n}\sum_{i=1}^n T_{i,n}^2 - \overline{T}_n^2.$$

The first requirement is to ensure almost sure convergence to 0 of both quantities $\left|\overline{T}_n - \mathbb{E}\left[T_{1,n}\right]\right|$ and $\left|\frac{1}{n}\sum_{i=1}^n T_{i,n}^2 - \overline{T}_n^2 - V_T\right|$. Under the assumption that $\sup_{n\in\mathbb{N}^*}\mathbb{E}\left[\left|T_{1,n}\right|^{4+\delta}\right] < +\infty$, observe that all the conditions of Theorem 2.2 in [84] are satisfied with $p = 1$. We can thus conclude that $\left|\overline{T}_n - \mathbb{E}\left[T_{1,n}\right]\right| \xrightarrow{a.s.} 0$ and $\left|\frac{1}{n}\sum_{i=1}^n T_{i,n}^2 - \mathbb{E}\left[T_{1,n}^2\right]\right| \xrightarrow{a.s.} 0$. Now using the fact that

$$\left|\frac{1}{n}\sum_{i=1}^n T_{i,n}^2 - \overline{T}_n^2 - V_T\right| \le \left|\frac{1}{n}\sum_{i=1}^n T_{i,n}^2 - \mathbb{E}\left[T_{1,n}^2\right]\right| + \left|\overline{T}_n - \mathbb{E}\left[T_{1,n}\right]\right|^2$$
$$+ 2\left|\mathbb{E}\left[T_{1,n}\right]\left(\overline{T}_n - \mathbb{E}\left[T_{1,n}\right]\right)\right| + \left|\mathbb{E}\left[T_{1,n}^2\right] - \mathbb{E}\left[T_{1,n}\right]^2 - V_T\right|,$$

as well as $|\mathbb{E}[T_{1,n}]| = O(1)$ and $||\mathbb{V}[(\gamma_{X,n}X_{1,n}\,,\,\gamma_{Y,n}Y_{1,n})] - V|| = o(1)$, to conclude that $\left|\frac{1}{n}\sum_{i=1}^{n} T_{i,n}^2 - \overline{T}_n^2 - V_T\right| \xrightarrow{a.s.} 0$.

The second requirement is to check the Lindeberg condition for the bootstrap which writes

$$\mathbb{E}\left[\left|T_{1,n}^{(1)}\right|^2 \mathbb{1}\left\{\left|T_{1,n}^{(1)}\right|^2 > \epsilon\sqrt{n}\right\} \mid (X_{i,n},Y_{i,n})_{i=1}^n\right] = \frac{1}{n}\sum_{i=1}^{n} |T_{i,n}|^2 \mathbb{1}\left\{|T_{i,n}|^2 > \epsilon\sqrt{n}\right\} \xrightarrow{a.s.} 0 \quad \forall \epsilon > 0.$$

Let $M : \epsilon \mapsto M(\epsilon)$ be some function of ϵ to be defined later that does not depend on n and satisfies $0 < M(\epsilon) < +\infty\ \forall \epsilon > 0$. For such a function, there exists for every $\epsilon > 0$, a n_ϵ such that for every $n > n_\epsilon$,

$$\frac{1}{n}\sum_{i=1}^{n} |T_{i,n}|^2 \mathbb{1}\left\{|T_{i,n}|^2 > \epsilon\sqrt{n}\right\} \leq \frac{1}{n}\sum_{i=1}^{n} |T_{i,n}|^2 \mathbb{1}\left\{|T_{i,n}|^2 > M(\epsilon)\right\} \ a.s.$$

By the triangle inequality,

$$\frac{1}{n}\sum_{i=1}^{n} |T_{i,n}|^2 \mathbb{1}\left\{|T_{i,n}|^2 > M(\epsilon)\right\} \leq \left|\frac{1}{n}\sum_{i=1}^{n} |T_{i,n}|^2 \mathbb{1}\left\{|T_{i,n}|^2 > M(\epsilon)\right\} - \mathbb{E}\left[|T_{1,n}|^2 \mathbb{1}\left\{|T_{1,n}|^2 > M(\epsilon)\right\}\right]\right| + \mathbb{E}\left[|T_{1,n}|^2 \mathbb{1}\left\{|T_{1,n}|^2 > M(\epsilon)\right\}\right].$$

The first term in the upper bound converges to 0 almost surely for every $\epsilon > 0$ under the assumption $\sup_{n\in\mathbb{N}^*} \mathbb{E}\left[|T_{1,n}|^{4+\delta}\right] < +\infty$ thanks to Theorem 2.2 in [84]. The second term in the upper bound can be bounded with the Cauchy-Schwarz and Markov inequalities

$$\mathbb{E}\left[|T_{1,n}|^2 \mathbb{1}\left\{|T_{1,n}|^2 > M(\epsilon)\right\}\right] \leq \frac{\sup_{n\in\mathbb{N}^*}\sqrt{\mathbb{E}\left[|T_{1,n}|^4\right]\mathbb{E}\left[|T_{1,n}|\right]}}{\sqrt{M(\epsilon)}}.$$

Picking $M(\epsilon) = \epsilon^{-1}\sup_{n\in\mathbb{N}^*} \mathbb{E}\left[|T_{1,n}|^4\right]\mathbb{E}\left[|T_{1,n}|\right]$, we get that for every $\epsilon > 0$

$$\limsup_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} |T_{i,n}|^2 \mathbb{1}\left\{|T_{i,n}|^2 > \epsilon\sqrt{n}\right\} \leq \epsilon \quad a.s.$$

Letting ϵ go to 0, we see that the Lindeberg condition is satisfied. This entails that (5.5) is satisfied.

Arguments underpinning the Cramer-Wold device are valid as well so that we can claim that for every $t \in \mathbb{R}^2$

$$\left|\mathbb{E}\left[e^{it'S_n^{(1)}} \mid (X_{i,n},Y_{i,n})_{i=1}^n\right] - e^{-t'Vt/2}\right| \xrightarrow{a.s.} 0, \tag{5.6}$$

where $S_n^{(1)} := \sqrt{n}\left(\gamma_{X,n}\left(\overline{X}_n^{(1)} - \overline{X}_n\right), \gamma_{Y,n}\left(\overline{Y}_n^{(1)} - \overline{Y}_n\right)\right)$.

Let Ω be the set of probability one on which (5.6) occurs. For every $\omega \in \Omega$,

$$\left(P_{S_n^{(1)}|(X_{i,n},Y_{i,n})_{i=1}^n = (X_{i,n}(\omega),Y_{i,n}(\omega))_{i=1}^n}\right)_{n\in\mathbb{N}^*}$$

is a sequence of nonrandom probability measures for which all weak convergence criteria are equivalent. In particular, for every $\omega \in \Omega$, the validity of the Lévy criterion due to (5.6) ensures that

$$d_{BL}\left(P_{S_n^{(1)}|(X_{i,n},Y_{i,n})_{i=1}^n = (X_{i,n}(\omega),Y_{i,n}(\omega))_{i=1}^n}, \mathcal{N}(0,V)\right) = o(1).$$

This is enough to conclude.

□

5.9.2.2 Proof of Lemma 5.5

The vector $\big(W_n^+, W_n^{(1)+}, W_n^{(2)+}\big)$ converges almost surely to $(0,0,0)$. As a consequence, there exists an event $\tilde{\Omega}^1$ of probability 1 such that $\forall \omega \in \tilde{\Omega}^1$, $\big(W_n^+(\omega),\, W_n^{(1)+}(\omega), W_n^{(2)+}(\omega)\big) = (0,0,0)$ for n large enough. As $\big(h_{Y,n}^+, h_{Y,n}^{(1)+}, h_{Y,n}^{(2)+}\big)$ converges almost surely to a continuous vector, there exists an event $\tilde{\Omega}^2$ of probability 1 such that $\forall \omega \in \tilde{\Omega}^2$, the components of $\big(h_{Y,n}^+(\omega), h_{Y,n}^{(1)+}(\omega), h_{Y,n}^{(2)+}(\omega)\big)$ are all non-zero for n large enough. We finally define $\tilde{\Omega} := \tilde{\Omega}^1 \cap \tilde{\Omega}^2$, which is of probability 1 and satisfies the stated conditions. □

5.9.3 Proof of Example 5.3

We have

$$\begin{aligned}
\mathbb{P}(\overline{Y}_n^{(1)} = 0) &= \mathbb{E}\Big[\mathbb{P}\big(\overline{Y}_n^{(1)} = 0 \,\big|\, (Y_{i,n})_{i=1}^n\big)\Big] \\
&= \mathbb{E}\Big[\mathbb{P}\big(Y_{1,n}^{(1)} = 0, \dots, Y_{n,n}^{(1)} = 0 \,\big|\, (Y_{i,n})_{i=1}^n\big)\Big] \\
&= \mathbb{E}\Big[\mathbb{P}\big(Y_{1,n}^{(1)} = 0 \,\big|\, (Y_{i,n})_{i=1}^n\big)^n\Big] = \mathbb{E}\Big[\big(S_n/n)\big)^n\Big],
\end{aligned}$$

where $S_n := \sum_{i=1}^n (1 - Y_{i,n}) \sim Bin(n, 1-p_n)$. Therefore, for any $x > 0$,

$$\begin{aligned}
\mathbb{P}(\overline{Y}_n^{(1)} = 0) &= \sum_{k=1}^{n} (k/n)^n \mathbb{P}\big[S_n = k\big] \\
&\le \sum_{k=1}^{\lfloor n(1-p_n)+x \rfloor} (k/n)^n \mathbb{P}\big[S_n = k\big] + \mathbb{P}\big[S_n \ge n(1-p_n) + x\big] \\
&\le \left(\frac{n(1-p_n)+x}{n}\right)^n + \mathbb{P}\big[S_n \ge n(1-p_n) + x\big] \\
&\le \big(1 - p_n + x/n\big)^n + \mathbb{P}\big[S_n - n(1-p_n) \ge x\big].
\end{aligned}$$

Let $\tilde{S}_n := \big(S_n - n(1-p_n)\big)/\sqrt{np_n(1-p_n)} = O_P(1)$ be the renormalized version of S_n and choose $x = n^a\sqrt{np_n(1-p_n)}$ for $a = (1-b)/3 > 0$. Then

$$\begin{aligned}
\mathbb{P}(\overline{Y}_n^{(1)} = 0) &\le \big(1 - p_n + n^a\sqrt{p_n(1-p_n)/n}\big)^n + \mathbb{P}\big[\tilde{S}_n \ge n^a\big] \\
&\le \exp\Big(n \ln\big(1 - p_n + n^a\sqrt{p_n(1-p_n)/n} + o(p_n)\big)\Big) + o(1) \\
&\le \exp\Big(n\big(n^{a-b/2-1/2} - n^{-b} + o(n^{-b})\big)\Big) + o(1) \\
&\le \exp\Big(n^{1/3-b/3-b/2+1/2} - n^{-b+1} + o(n^{-b+1})\Big) + o(1) \\
&\le \exp\Big(n^{(1-b)5/6} - n^{1-b} + o(n^{-b+1})\Big) + o(1) \\
&\le \exp\big(-n^{1-b}\big) + o(1) = o(1),
\end{aligned}$$

which completes the proof. □

5.9.4 Proof of Theorem 5.3

We fix arbitrary $n \in \mathbb{N}^*$ and $\varepsilon \in \mathbb{R}_+^*$. Combining the triangle inequality, the bound $|\overline{X}_n| \le |\overline{X}_n - \mathbb{E}[X_{1,n}]| + |\mathbb{E}[X_{1,n}]|$ and Assumptions 5.1 to 5.3, we get

$$\left|\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}[X_{1,n}]}{\mathbb{E}[Y_{1,n}]}\right| \le |\overline{X}_n| \cdot \left|\frac{1}{\overline{Y}_n} - \frac{1}{\mathbb{E}[Y_{1,n}]}\right| + \frac{1}{\mathbb{E}[Y_{1,n}]}\left|\overline{X}_n - \mathbb{E}[X_{1,n}]\right|$$

$$\leq \frac{\left(\left|\overline{X}_n - \mathbb{E}\left[X_{1,n}\right]\right| + \sqrt{u_{X,n}}\right)\left|\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right]\right|}{a_{Y,n} l_{Y,n}} + \frac{\left|\overline{X}_n - \mathbb{E}\left[X_{1,n}\right]\right|}{l_{Y,n}}.$$

Consequently, the event considered in Theorem 5.3 is included in the event

$$\frac{\left(\left|\overline{X}_n - \mathbb{E}\left[X_{1,n}\right]\right| + \sqrt{u_{X,n}}\right)\left|\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right]\right|}{a_{Y,n} l_{Y,n}} + \frac{\left|\overline{X}_n - \mathbb{E}\left[X_{1,n}\right]\right|}{l_{Y,n}} > \frac{\left(\varepsilon + \sqrt{u_{X,n}}\right)\varepsilon}{a_{Y,n} l_{Y,n}} + \frac{\varepsilon}{l_{Y,n}}. \tag{5.7}$$

If both $|\overline{X}_n - \mathbb{E}\left[X_{1,n}\right]|$ and $|\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right]|$ are inferior or equal to ε, event (5.7) cannot happen. By contraposition, we obtain:

$$\begin{aligned}
&\mathbb{P}\left(\frac{\left(\left|\overline{X}_n - \mathbb{E}\left[X_{1,n}\right]\right| + \sqrt{u_{X,n}}\right)\left|\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right]\right|}{a_{Y,n} l_{Y,n}} + \frac{\left|\overline{X}_n - \mathbb{E}\left[X_{1,n}\right]\right|}{l_{Y,n}} > \frac{\left(\varepsilon + \sqrt{u_{X,n}}\right)\varepsilon}{a_{Y,n} l_{Y,n}} + \frac{\varepsilon}{l_{Y,n}}\right) \\
&\leq \mathbb{P}\left(\left\{\left|\overline{X}_n - \mathbb{E}\left[X_{1,n}\right]\right| > \varepsilon\right\} \cup \left\{\left|\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right]\right| > \varepsilon\right\}\right) \\
&\leq \mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}\left[X_{1,n}\right]\right| > \varepsilon\right) + \mathbb{P}\left(\left|\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right]\right| > \varepsilon\right),
\end{aligned}$$

where we use the union bound for the last inequality. The first conclusion follows from using twice Bienaymé-Chebyshev's inequality applied to the variables $\overline{X}_n$ and $\overline{Y}_n$ and the fact that under Assumptions 5.1 and 5.2 and Jensen's inequality, $\mathbb{V}\left[X_{1,n}\right] \leq u_{X,n}$ and $\mathbb{V}\left[Y_{1,n}\right] \leq u_{Y,n} - l_{Y,n}^2$. The second conclusion follows from solving $(u_{X,n} + u_{Y,n} - l_{Y,n}^2)/(n\varepsilon^2) = \alpha$.

□

5.9.5 Proof of Theorem 5.4

We start by introducing and proving an intermediate lemma that is also used to prove Theorem 5.9. For a random variable U, $\varepsilon > 0$, and $\tilde{\varepsilon} \in (0,1)$ we define the following events:

$$A_\varepsilon^U := \left\{\left|\overline{U}_n - \mathbb{E}[U]\right| \leq \varepsilon\right\}, \text{ and } \tilde{A}_{\tilde{\varepsilon}}^U := \left\{\left|\overline{U}_n - \mathbb{E}[U]\right| \leq \tilde{\varepsilon}\left|\mathbb{E}[U]\right|\right\}.$$

Lemma 5.6. *Assume that Assumption 5.1 holds. Then for every $n \in \mathbb{N}^*$, $\varepsilon > 0$ and $\tilde{\varepsilon} \in (0,1)$, we have*

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}\left[X_{1,n}\right]}{\mathbb{E}\left[Y_{1,n}\right]}\right| > \left(\frac{\left(\left|\mathbb{E}\left[X_{1,n}\right]\right| + \varepsilon\right)\tilde{\varepsilon}}{(1-\tilde{\varepsilon})^2} + \varepsilon\right)\frac{1}{\left|\mathbb{E}\left[Y_{1,n}\right]\right|}\right) \\
&\leq 1 - \mathbb{P}\left(A_\varepsilon^{X_{1,n}}\right) + 1 - \mathbb{P}\left(\tilde{A}_{\tilde{\varepsilon}}^{Y_{1,n}}\right).
\end{aligned}$$

We fix arbitrary $n \in \mathbb{N}^*$, $\varepsilon > 0$ and $\tilde{\varepsilon} \in (0,1)$. By Lemma 5.6, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}\left[X_{1,n}\right]}{\mathbb{E}\left[Y_{1,n}\right]}\right| > \left(\frac{\left(\left|\mathbb{E}\left[X_{1,n}\right]\right| + \varepsilon\right)\tilde{\varepsilon}}{(1-\tilde{\varepsilon})^2} + \varepsilon\right)\frac{1}{\left|\mathbb{E}\left[Y_{1,n}\right]\right|}\right) \\
&\leq 1 - \mathbb{P}\Big(\left|\overline{X}_n - \mathbb{E}\left[X_{1,n}\right]\right| \leq \varepsilon\Big) + 1 - \mathbb{P}\Big(\left|\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right]\right| \leq \tilde{\varepsilon}\left|\mathbb{E}\left[Y_{1,n}\right]\right|\Big).
\end{aligned}$$

Using Jensen's inequality and Assumption 5.2, we have $|\mathbb{E}\left[X_{1,n}\right]| \leq (u_{X,n})^{1/2}$, and Assumption 5.1 entails $1/|\mathbb{E}\left[Y_{1,n}\right]| \leq 1/l_{Y,n}$. Consequently, we get

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}\left[X_{1,n}\right]}{\mathbb{E}\left[Y_{1,n}\right]}\right| > \left(\frac{\left(\sqrt{u_{X,n}} + \varepsilon\right)\tilde{\varepsilon}}{(1-\tilde{\varepsilon})^2} + \varepsilon\right)\frac{1}{l_{Y,n}}\right) \\
&\leq 1 - \mathbb{P}\Big(\left|\overline{X}_n - \mathbb{E}\left[X_{1,n}\right]\right| \leq \varepsilon\Big) + 1 - \mathbb{P}\Big(\left|\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right]\right| \leq \tilde{\varepsilon}\left|\mathbb{E}\left[Y_{1,n}\right]\right|\Big).
\end{aligned}$$

Using Bienaymé-Chebyshev's inequality twice gives the bounds

$$1 - \mathbb{P}\Big(\big|\overline{X}_n - \mathbb{E}\left[X_{1,n}\right]\big| \leq \varepsilon\Big) \leq \frac{\mathbb{V}\left[X_{1,n}\right]}{n\varepsilon^2}$$
$$1 - \mathbb{P}\Big(\big|\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right]\big| \leq \tilde{\varepsilon}\big|\mathbb{E}\left[Y_{1,n}\right]\big|\Big) \leq \frac{\mathbb{V}\left[Y_{1,n}\right]}{n\tilde{\varepsilon}^2\left(\mathbb{E}\left[Y_{1,n}\right]\right)^2}.$$

For the numerator, $\mathbb{V}\left[X_{1,n}\right] = \mathbb{E}\left[X_{1,n}^2\right] - \left(\mathbb{E}\left[X_{1,n}\right]\right)^2 \leq \mathbb{E}\left[X_{1,n}^2\right] \leq u_{X,n}$ using Assumption 5.2. For the denominator, Assumption 5.1 immediately entails that $1/l_{Y,n}^2$ is an upper bound on $1/\left(\mathbb{E}\left[Y_{1,n}\right]\right)^2$ and $l_{Y,n}^2$ a lower bound on $\left(\mathbb{E}\left[Y_{1,n}\right]\right)^2$. Therefore

$$\frac{\mathbb{V}\left[Y_{1,n}\right]}{n\tilde{\varepsilon}^2\left(\mathbb{E}\left[Y_{1,n}\right]\right)^2} \leq \frac{\mathbb{E}\left[Y_{1,n}^2\right] - {l_{Y,n}}^2}{n\tilde{\varepsilon}^2{l_{Y,n}}^2} \leq \frac{u_{Y,n} - {l_{Y,n}}^2}{n\tilde{\varepsilon}^2{l_{Y,n}}^2},$$

where the second inequality uses Assumption 5.2.

Combining the two bounds yields the following upper bound on the probability considered in Theorem 5.4

$$\frac{u_{X,n}}{n\varepsilon^2} + \frac{u_{Y,n} - {l_{Y,n}}^2}{n\tilde{\varepsilon}^2{l_{Y,n}}^2}, \tag{5.8}$$

as claimed.

For the second part of Theorem 5.4, for a fixed α, we equalize each of the two terms in (5.8) to $\alpha/2$ and solve for ε and $\tilde{\varepsilon}$, which yields:

$$\varepsilon^2 = \frac{2u_{X,n}}{n\alpha} \text{ and } \tilde{\varepsilon}^2 = \frac{2\left(u_{Y,n} - {l_{Y,n}}^2\right)}{n\alpha l_{Y,n}^2}.$$

The bound $\overline{\alpha}_n$ comes from the fact that $\tilde{\varepsilon}$ needs to be smaller than 1.

□

5.9.5.1 Proof of Lemma 5.6

We fix arbitrary $\varepsilon > 0$ and $\tilde{\varepsilon} \in (0,1)$. Without loss of generality, we can assume that $\mathbb{E}\left[Y_{1,n}\right] > 0$ and $\mathbb{E}\left[X_{1,n}\right] \geq 0$.

First, using the union bound, note that the event $A_\varepsilon^{X_{1,n}} \cap \tilde{A}_{\tilde{\varepsilon}}^{Y_{1,n}}$ holds with a probability bigger than $\mathbb{P}\left(A_\varepsilon^{X_{1,n}}\right) + \mathbb{P}\left(\tilde{A}_{\tilde{\varepsilon}}^{Y_{1,n}}\right) - 1$. Hence, its complement is of probability lower than $1 - \mathbb{P}\left(A_\varepsilon^{X_{1,n}}\right) + 1 - \mathbb{P}\left(\tilde{A}_{\tilde{\varepsilon}}^{Y_{1,n}}\right)$.

Second, we show that the event considered in Lemma 5.6 is included in the complement of $A_\varepsilon^{X_{1,n}} \cap \tilde{A}_{\tilde{\varepsilon}}^{Y_{1,n}}$, which concludes the proof. To do so, we reason by contraposition and do the following computations on the event $A_\varepsilon^{X_{1,n}} \cap \tilde{A}_{\tilde{\varepsilon}}^{Y_{1,n}}$.

By the triangle inequality, we get

$$\left|\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}\left[X_{1,n}\right]}{\mathbb{E}\left[Y_{1,n}\right]}\right| \leq |\overline{X}_n| \cdot \left|\frac{1}{\overline{Y}_n} - \frac{1}{\mathbb{E}\left[Y_{1,n}\right]}\right| + \frac{1}{\mathbb{E}\left[Y_{1,n}\right]}\left|\overline{X}_n - \mathbb{E}\left[X_{1,n}\right]\right|.$$

We now bound the first term using the mean value theorem applied to the function $f(x) := 1/(x + \mathbb{E}\left[Y_{1,n}\right])$

$$\left|\frac{1}{\overline{Y}_n} - \frac{1}{\mathbb{E}\left[Y_{1,n}\right]}\right| = \left|f(\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right]) - f(0)\right| \leq \frac{|\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right]|}{(1-\tilde{\varepsilon})^2\mathbb{E}\left[Y_{1,n}\right]^2}$$
$$\leq \frac{\tilde{\varepsilon}\mathbb{E}\left[Y_{1,n}\right]}{(1-\tilde{\varepsilon})^2\mathbb{E}\left[Y_{1,n}\right]^2},$$

where the first inequality uses the following observation: on the event $\tilde{A}_{\tilde{\varepsilon}}^{Y_{1,n}}$, a lower bound on $|x + \mathbb{E}[Y_{1,n}]|$ with x varying between 0 and $\overline{Y}_n - \mathbb{E}[Y_{1,n}]$ is $(1-\tilde{\varepsilon})|\mathbb{E}[Y_{1,n}]|$. Therefore, on $A_{\varepsilon}^{X_{1,n}} \cap \tilde{A}_{\tilde{\varepsilon}}^{Y_{1,n}}$,

$$\begin{aligned}\left|\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}[X_{1,n}]}{\mathbb{E}[Y_{1,n}]}\right| &\leq |\overline{X}_n| \cdot \frac{\tilde{\varepsilon}\mathbb{E}[Y_{1,n}]}{(1-\tilde{\varepsilon})^2\mathbb{E}[Y_{1,n}]^2} + \frac{\varepsilon}{\mathbb{E}[Y_{1,n}]} \\ &\leq \left(|\mathbb{E}[X_{1,n}]| + |\overline{X}_n - \mathbb{E}[X_{1,n}]|\right)\frac{\tilde{\varepsilon}}{(1-\tilde{\varepsilon})^2\mathbb{E}[Y_{1,n}]} + \frac{\varepsilon}{\mathbb{E}[Y_{1,n}]} \\ &\leq \frac{\left(|\mathbb{E}[X_{1,n}]| + \varepsilon\right)\tilde{\varepsilon}}{(1-\tilde{\varepsilon})^2\mathbb{E}[Y_{1,n}]} + \frac{\varepsilon}{\mathbb{E}[Y_{1,n}]},\end{aligned}$$

where we use the triangle inequality to get the second line. It is indeed the complement of the event considered in the statement of Lemma 5.6.

□

5.9.6 Proof of Theorem 5.5

This theorem relies crucially on the following lemma.

Lemma 5.7. *For each ξ in the interval $\left(0, 1 \wedge \left(u_{Y,n}/l_{Y,n}^2 - 1\right)\right)$, there exists a distribution $P_{n,\xi} \in \mathcal{P}$ such that $\mathbb{P}\left(\overline{Y}_n = 0\right) \geq \tilde{\alpha}_n(\xi)$, where $\tilde{\alpha}_n(\xi) := \left(1 - (1+\xi)l_{Y,n}^2/u_{Y,n}\right)^n$.*

Note that the interval $\left(0, 1 \wedge \left(u_{Y,n}/l_{Y,n}^2 - 1\right)\right)$ is not empty since we have assumed $u_{Y,n}/l_{Y,n}^2 > 1$.

By Lemma 5.7, for every $\xi < 1 \wedge \left(u_{Y,n}/l_{Y,n}^2 - 1\right)$, there exists a distribution $P_{n,\xi}$ such that $\mathbb{P}\left(\overline{Y}_n = 0\right) \geq \tilde{\alpha}_n(\xi)$. Taking the supremum over ξ, we deduce that

$$\sup_{P_n \in \mathcal{P}} \mathbb{P}\left(\overline{Y}_n = 0\right) \geq \sup_{\xi} \tilde{\alpha}_n(\xi) = \underline{\alpha}_n.$$

Using the assumption that I_n is undefined whenever $\overline{Y}_n = 0$, we deduce that $\mathbb{P}\big(I_n\, undefined\big) \geq \underline{\alpha}_n$.

□

5.9.6.1 Proof of Lemma 5.7

We consider the following distribution on $\mathbb{R}$

$$P_{n,l_{Y,n},u_{Y,n},c,\xi} := \left(\frac{c}{n}\right)^{1/n}\delta_{\{0\}} + \frac{1}{2}\left(1 - \left(\frac{c}{n}\right)^{1/n}\right)\delta_{\{y_{c-}\}} + \frac{1}{2}\left(1 - \left(\frac{c}{n}\right)^{1/n}\right)\delta_{\{y_{c+}\}},$$

where $c \in (0,n)$ is some constant to be chosen later, $y_{c-} := l_{Y,n}(1-\sqrt{\xi})/(1-(c/n)^{1/n})$ and $y_{c+} := l_{Y,n}(1+\sqrt{\xi})/(1-(c/n)^{1/n})$. Let $Y_{1,n} \sim P_{n,l_{Y,n},u_{Y,n},c,\xi_n}$. Observe that $\mathbb{E}[Y_{1,n}] = l_{Y,n}$ and $\mathbb{E}\left[Y_n^2\right] = l_{Y,n}^2(1+\xi_n)/\left(1-(c/n)^{1/n}\right)$. With the choice

$$c = c_n := n\left(1 - \frac{l_{Y,n}^2}{u_{Y,n}}(1+\xi)\right)^n,$$

we have $\mathbb{E}\left[Y_n^2\right] = u_{Y,n}$. Note that $C_{n,\alpha}$ is strictly positive, because $1 - \frac{l_{Y,n}^2}{u_{Y,n}}(1+\xi_n)$ is positive. This is equivalent to $u_{Y,n}/l_{Y,n}^2 > 1 + \xi_n$, which is true by assumption.

Consider now the following product measure on $\mathbb{R}^2$ defined by $P_n := \delta_{\left\{\sqrt{u_{X,n}}\right\}} \otimes P_{n,l_{Y,n},u_{Y,n},c_n,\xi}$. Let $(X_{i,n}, Y_{i,n})_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_n$. These random vectors satisfy $\mathbb{E}\left[X_n^2\right] = u_{X,n}$, $\mathbb{E}[Y_{1,n}] = l_{Y,n}$ and $\mathbb{E}\left[Y_n^2\right] = u_{Y,n}$. The next step is to build a lower bound on the event $\{\overline{Y}_n = 0\}$.

The assumption that $(X_{i,n}, Y_{i,n})_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_n$ and the construction of $P_{n,l_{Y,n},u_{Y,n},c_n,\xi}$ imply that

$$\mathbb{P}\left(\overline{Y}_n = 0\right) = \frac{c_n}{n} = \left(1 - \frac{l_{Y,n}^2}{u_{Y,n}}(1+\xi)\right)^n = \tilde{\alpha}_n(\xi).$$

□

5.9.7 Proof of Theorem 5.6

To prove Theorem 5.6, we need the following lemma.

Lemma 5.8. *For every integer* $n \geq 7$ *and every* $x \in (0,1)$, $x\left(1-x/n\right)^{n-1} \geq x/3$.

We start using arguments developed in the proof of [34][Proposition 6.2]. We detail those for the sake of clarity. For every $n \in \mathbb{N}^*$ and $\eta > \sqrt{u_{X,n}}/n$, let us define the following distribution on $\mathbb{R}$, which will be used for the variable in the numerator[18]:

$$P_{n,u_{X,n},\eta} := \frac{u_{X,n}}{2n^2\eta^2}\delta_{\{-n\eta\}} + \left(1 - \frac{u_{X,n}}{n^2\eta^2}\right)\delta_{\{0\}} + \frac{u_{X,n}}{2n^2\eta^2}\delta_{\{n\eta\}}.$$

This distribution is symmetric, centered and has variance $u_{X,n}$. As shown in [34], every i.i.d. sample $(X_{i,n})_{i=1}^n$ drawn from $P_{n,u_{X,n},\eta}$ satisfies

$$\begin{aligned}&\mathbb{P}\left(\overline{X}_n \leq -\eta\right) = \mathbb{P}\left(\overline{X}_n \geq \eta\right) \geq \mathbb{P}\left(\overline{X}_n = \eta\right)\\ &\geq \sum_{i=1}^n \mathbb{P}\left(X_{i,n} = n\eta, X_{j,n} = 0,\ \forall j \neq i\right) = \frac{u_{X,n}}{2n\eta^2}\left(1 - \frac{u_{X,n}}{\eta^2 n^2}\right)^{n-1}.\end{aligned}$$

Note further that for every integer $n \geq 2$, $\mathbb{P}\left(\overline{X}_n \geq \eta\right) \geq \mathbb{P}\left(\overline{X}_n = \eta\right)$ becomes a strict inequality strict and for every $\xi \in (0,1)$ $\left\{\left|\overline{X}_n\right| \geq \eta\right\} \subseteq \left\{\left|\overline{X}_n\right| > \xi\eta\right\}$. As a result, if $(X_{i,n})_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{n,u_{X,n},\eta}$, for every $\eta > 0$, we have

$$\mathbb{P}\left(\left|\overline{X}_n\right| > \xi\eta\right) > \frac{u_{X,n}}{n\eta^2}\left(1 - \frac{u_{X,n}}{\eta^2 n^2}\right)^{n-1}. \tag{5.9}$$

The following steps do not show up in [34] since they are specific to controlling ratios of expectations and sample averages. For every $n \in \mathbb{N}^*$, let us define the following distribution on $\mathbb{R}$, which will be used for the variable in the denominator

$$P_{n,l_{Y,n},u_{Y,n}} := \frac{1}{2}\delta_{\left\{l_{Y,n} - \sqrt{u_{Y,n} - l_{Y,n}^2}\right\}} + \frac{1}{2}\delta_{\left\{l_{Y,n} + \sqrt{u_{Y,n} - l_{Y,n}^2}\right\}}.$$

Let $(X_{i,n}, Y_{i,n})_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_n := P_{n,u_{X,n},\eta} \otimes P_{n,l_{Y,n},u_{Y,n}}$. Observe that $\mathbb{E}\left[Y_{1,n}\right] = l_{Y,n}$ and $\mathbb{E}\left[Y_n^2\right] = u_{Y,n}$. Furthermore, $\left|\overline{Y}_n\right| \leq l_{Y,n} + \sqrt{u_{Y,n} - l_{Y,n}^2}$ almost surely. This implies that for every $\eta > 0$ and $\xi \in (0,1)$, the following holds

$$\left\{\left|\overline{X}_n\right| > \left(l_{Y,n} + \sqrt{u_{Y,n} - l_{Y,n}^2}\right)\xi\eta\right\} \subseteq \left\{\left|\frac{\overline{X}_n}{\overline{Y}_n}\right| > \xi\eta\right\}.$$

For fixed $n \geq 7$ and $\alpha \in \left(0, 1 \wedge n/\left(l_{Y,n} + \sqrt{u_{Y,n} - l_{Y,n}^2}\right)^2\right)$, we choose $\eta = \eta(\alpha) = \sqrt{v_n/3n\alpha}$. Combining the above inclusion with (5.9), and Lemma 5.8 (with the choice $x = 3\alpha$), we conclude that there exists a distribution on $\mathbb{R}^2$, namely P_n, that fulfills Assumptions 5.1 and 5.2 such that

$$\mathbb{P}\left(\left|\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}\left[X_{1,n}\right]}{\mathbb{E}\left[Y_{1,n}\right]}\right| > \xi\sqrt{\frac{v_n}{3n\alpha}}\right) > \alpha,$$

which completes the proof.

□

[18] The notation δ denotes the Dirac distribution.

5.9.7.1 Proof of Lemma 5.8

Under our assumptions on n and x, $\ln(1-x/n)$ is well-defined. Using Taylor-Lagrange formula on the function $[0,x] \ni t \mapsto \ln(1-t/n)$ yields:

$$\begin{aligned}\left(1-\frac{x}{n}\right)^{n-1} &= \exp\left((n-1)\ln\left(1-\frac{x}{n}\right)\right)\\ &= \exp\left(-(n-1)\left(\frac{x}{n}+\frac{1}{2\left(1-\tau x/n\right)^2}\frac{x^2}{n^2}\right)\right)\end{aligned}$$

for some $\tau \in (0,1)$. Using the fact that $\frac{n-1}{n} \leq 1$, $x \leq 1$ and $\frac{1}{2(1-\tau x/n)^2} \leq \frac{1}{2(1-n^{-1})^2}$, we get that under our assumptions $\left(1-\frac{x}{n}\right)^{n-1} \geq \exp\left(-\left(1+\frac{1}{2n(1-n^{-1})^2}\right)\right)$. This bound is actually valid for every $x \in (0,1)$ and every $n \in \mathbb{N}^*$. The computation of $\exp\left(-\left(1+\frac{1}{2n(1-n^{-1})^2}\right)\right)$ shows that the latter is larger than $1/4$ whenever $n \geq 3$ and larger than $1/3$ whenever $n \geq 7$.

□

5.10 Adapted results for "Hoeffding" framework

Assumption 5.4. *For every $n \in \mathbb{N}^*$, there exist finite constants $a_{X,n}$, $b_{X,n}$, $a_{Y,n}$, $b_{Y,n}$ and $l_{Y,n}$ such that $X_{1,n}$ (respectively $Y_{1,n}$) lies $P_{X,Y,n}$-almost surely in the interval $[a_{X,n},b_{X,n}]$ (resp. $[a_{Y,n},b_{Y,n}]$) and $|\mathbb{E}[Y_{1,n}]| \geq l_{Y,n}$.*

The support of $X_{1,n}$ and $Y_{1,n}$ is allowed to change with n, even though in many examples of interest, the former can be chosen independent from n. Assumptions 5.1 and 5.4 together correspond to the *Hoeffding case* because under these two assumptions, we can use the Hoeffding inequality to build nonasymptotic CIs.

5.10.1 Concentration inequality in an easy case: the support of the denominator is well-separated from 0

Assumption 5.5. *For every $n \in \mathbb{N}^*$, the lower bound $a_{Y,n}$ is strictly positive.*

Theorem 5.8. *Let $u_{X,n} := (b_{X,n}-a_{X,n})^2$ and $u_{Y,n} := (b_{Y,n}-a_{Y,n})^2$. Under Assumptions 5.1, 5.4 and 5.5, we have for every $n \in \mathbb{N}^*$ and $\varepsilon \in \mathbb{R}_+^*$*

$$\begin{aligned}\sup_{P\in\mathcal{P}} \mathbb{P}_{P^{\otimes n}}&\left(\left|\frac{\overline{X}_n}{\overline{Y}_n}-\frac{\mathbb{E}[X_{1,n}]}{\mathbb{E}[Y_{1,n}]}\right| > \frac{\varepsilon}{l_{Y,n}}\left\{1+\frac{1}{a_{Y,n}}\left(|a_{X,n}|\vee|b_{X,n}|+\varepsilon\right)\right\}\right)\\ &\leq 4\exp\left(-\frac{2n\varepsilon^2}{u_{X,n}\vee u_{Y,n}}\right).\end{aligned}$$

As a consequence, $\inf_{P\in\mathcal{P}} \mathbb{P}_{P^{\otimes n}}\left(\mathbb{E}[X_{1,n}]/\mathbb{E}[Y_{1,n}] \in \left[\overline{X}_n/\overline{Y}_n \pm t\right]\right) \geq 1-\alpha$*, with the following choice for t:*

$$\sqrt{\frac{(u_{X,n}\vee u_{Y,n})\ln(4/\alpha)}{2n{l_{Y,n}}^2}}\left(1+\frac{1}{a_{Y,n}}\left(|a_{X,n}|\vee|b_{X,n}|+\sqrt{\frac{(u_{X,n}\vee u_{Y,n})\ln(4/\alpha)}{2n}}\right)\right),$$

for every $\alpha \in (0,1)$.

The theorem shows that it is possible to construct nonasymptotic CIs for ratios of expectations at every confidence level that are almost surely bounded. However, it requires the additional Assumption 5.5, that in particular does not allow for binary $\{0,1\}$ random variables in the denominator which may

limit its applicability for various applications. In Section 5.10.2, we give an analogous result that only requires Assumptions 5.1 and 5.4 to hold, so that it encompasses the case of $\{0,1\}$-valued denominators. However, the cost to pay will be an upper bound on the achievable coverage of the confidence intervals.

5.10.2 Concentration inequality in the general case

We seek to build nontrivial nonasymptotic CIs under Assumptions 5.1 and 5.4 only. Under Assumption 5.1, $\mathbb{E}[Y_{1,n}] \neq 0$, so that there is no issue in considering the fraction $\mathbb{E}[X_{1,n}]/\mathbb{E}[Y_{1,n}]$. However, without Assumption 5.5, $\{\overline{Y}_n = 0\}$ has positive probability in general so that $\overline{X}_n/\overline{Y}_n$ is well-defined with probability less than one and undefined else. Note that when $P_{Y,n}$ is continuous wrt to Lebesgue's measure, there is no issue in defining $\overline{X}_n/\overline{Y}_n$ anymore since the event $\{\overline{Y}_n = 0\}$ has probability zero. This is not an easier case to establish concentration inequalities though, since without more restrictions, $\overline{Y}_n$ can still be arbitrarily close to 0 with positive probability.

Theorem 5.9. *Assume that Assumptions 5.1 and 5.4 hold. For every $n \in \mathbb{N}^*$, $\varepsilon > 0, \tilde{\varepsilon} \in (0,1)$, we have*

$$\sup_{P\in\mathcal{P}} \mathbb{P}_{P^{\otimes n}}\left(\left|\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}[X_{1,n}]}{\mathbb{E}[Y_{1,n}]}\right| > \left(\frac{(|a_{X,n}| \vee |b_{X,n}| + \varepsilon)\tilde{\varepsilon}}{(1-\tilde{\varepsilon})^2} + \varepsilon\right)\frac{1}{l_{Y,n}}\right)$$
$$\leq 2\exp(-n\varepsilon^2\gamma(X_{1,n})) + 2\exp(-n\tilde{\varepsilon}^2\gamma(Y_{1,n})),$$

where $\gamma(X_{1,n}) = 2/(b_{X,n}-a_{X,n})^2$ and $\gamma(Y_{1,n}) = 2l_{Y,n}^2/(b_{Y,n}-a_{Y,n})^2$.

As a consequence, $\inf_{P\in\mathcal{P}} \mathbb{P}_{P^{\otimes n}}\left(\mathbb{E}[X_{1,n}]/\mathbb{E}[Y_{1,n}] \in \left[\overline{X}_n/\overline{Y}_n \pm t\right]\right) \geq 1-\alpha$, with the choice

$$t := \sqrt{\frac{\ln(4/\alpha)}{n\gamma(X_{1,n}) \wedge \gamma(Y_{1,n})}}\left(\frac{|a_{X,n}| \vee |b_{X,n}| + \sqrt{\ln(4/\alpha)/(n\gamma(X_{1,n}))}}{\left(1-\sqrt{\ln(4/\alpha)/(n\gamma(Y_{1,n}))}\right)^2} + 1\right)\frac{1}{l_{Y,n}},$$

for every $\alpha > \overline{\alpha}_{n,H} := 4e^{-n\gamma(Y_{1,n})}$.[19]

This theorem is proved in Section 5.10.4. It states that when $l_{Y,n} > 0$, it is possible to build valid nonasymptotic CIs with finite length up to the confidence level $1-\overline{\alpha}_{n,H}$. This is a more positive result than [63] which claims that it is not possible to build nontrivial nonasymptotic CIs when $l_{Y,n}$ is taken equal to 0, no matter the confidence level. Note that Theorem 5.9 is not an impossibility theorem since it only claims that considering confidence levels smaller than $1-\overline{\alpha}_{n,H}$ is *sufficient* to build nontrivial CIs under Assumptions 5.1 and 5.4. The remaining question is to find out whether it is *necessary* to focus on confidence levels that do not exceed a certain threshold under Assumptions 5.1 and 5.4. We answer this in Section 5.10.3.

Theorem 5.9 has two other interesting consequences: for every confidence level up to $1-\overline{\alpha}_{n,H}$, a nonasymptotic CI of the form $\left[\overline{X}_n/\overline{Y}_n \pm \tilde{t}\right]$ with $\tilde{t} > t$ has good coverage but is too conservative. What is more, if the DGP does not depend on n (*i.e* in the standard i.i.d. set-up), for every fixed $\alpha > \overline{\alpha}_{n,H}$, the length of the confidence interval shrinks at the optimal rate $1/\sqrt{n}$.

5.10.3 An upper bound on testable confidence levels

Theorem 5.10. *For every $n \in \mathbb{N}^*$, and every $\alpha \in \left(0, \underline{\alpha}_{n,H}\right)$, where $\underline{\alpha}_{n,H} := \left(1 - l_{Y,n}/(b_{Y,n}-a_{Y,n})\right)^n$, if $(b_{Y,n}-a_{Y,n})/l_{Y,n} > 1$, there is no finite $t > 0$ such that $\left[\overline{X}_n/\overline{Y}_n \pm t\right]$ has coverage $1-\alpha$ over $\mathcal{P}_H$, where $\mathcal{P}_H$ is the class of all distributions satisfying Assumptions 5.1 and 5.4 for a fixed lower bound $l_{Y,n}$ and fixed lengths $b_{X,n}-a_{X,n}$ and $b_{Y,n}-a_{Y,n}$.*

[19] Equivalently, it means that for a given level α, the choice of t is valid for every integer $n > \overline{n}_{\alpha,H} := \ln(4/\alpha)/\gamma(Y_{1,n})$.

This theorem asserts that confidence intervals of the form $\left[\overline{X}_n/\overline{Y}_n \pm t\right]$ with coverage higher than $1-\underline{\alpha}_{n,H}$ under Assumptions 5.1 and 5.4 are not defined (or are of infinite length) with positive probability for at least one distribution in $\mathcal{P}_H$. The additional restriction $(b_{Y,n}-a_{Y,n})/l_{Y,n}>1$ is rather mild in practice: it is equivalent to $b_{Y,n}-a_{Y,n}>l_{Y,n}$ and is satisfied as soon as $a_{Y,n}\leq 0$ and $b_{Y,n}>l_{Y,n}>0$. This encompasses all DGPs where the denominator is $\{0,1\}$-valued and the probability that the denominator equals 1 is bounded from below by $l_{Y,n}\in(0,1)$.

Note that for Theorems 5.8 and 5.9, it is required to know not only the length $b_{X,n}-a_{X,n}$ but also the actual endpoints of the support, $a_{X,n}$ and $b_{X,n}$. On the contrary, Theorem 5.10 does not require the latter. In that respect, the class of Theorem 5.10 is larger than the one of the two preceding theorems.

5.10.4 Proof of Theorems 5.8 and 5.9

The proofs are identical to those of Theorems 5.3 and 5.4, except for the Bienaymé-Chebyshev inequality that has to be replaced with the Hoeffding inequality. The latter can be used under Assumption 5.4. Note also that $\mathbb{E}\left[X_{1,n}\right]$ is now bounded by $|a_{X,n}|\vee|b_{X,n}|$.

□

5.10.5 Proof of Theorem 5.10

We need the subsequent lemma.

Lemma 5.9. *For each ξ in the interval $\Big(0,1\wedge\big((b_{Y,n}-a_{Y,n})/l_{Y,n}-1\big)\Big)$, there exists a distribution $P_{n,\xi}\in\mathcal{P}_H$ such that $\mathbb{P}\left(\overline{Y}_n=0\right)\geq\tilde{\alpha}_{n,H}(\xi)$, where $\tilde{\alpha}_{n,H}(\xi):=\big(1-(1+\xi)l_{Y,n}/(b_{Y,n}-a_{Y,n})\big)^n$.*

Note that the interval $\Big(0,1\wedge\big((b_{Y,n}-a_{Y,n})/l_{Y,n}-1\big)\Big)$ is not empty since we have assumed $(b_{Y,n}-a_{Y,n})/l_{Y,n}>1$.

By Lemma 5.9, for every $\xi<1\wedge\big((b_{Y,n}-a_{Y,n})/l_{Y,n}-1\big)$, there exists a distribution $P_{n,\xi}\in\mathcal{P}_H$ satisfying Assumptions 5.1 and 5.4 such that $\mathbb{P}\left(\overline{Y}_n=0\right)\geq\tilde{\alpha}_{n,H}(\xi)$. Denote its marginal distributions by $P_{X,n,\xi}$ and $P_{Y,n,\xi}$. Therefore, $P_{n,\xi}$ satisfies Assumptions 5.1 and 5.4, and $\overline{X}_n/\overline{Y}_n$ is undefined with probability greater than $\tilde{\alpha}_{n,H}(\xi)$. Taking the supremum over ξ, we deduce that

$$\sup_{P_n\in\mathcal{P}_H}\mathbb{P}\left(\overline{Y}_n=0\right)\geq\sup_{\xi}\tilde{\alpha}_n(\xi)=\underline{\alpha}_{n,H}.$$

This means that the random interval $I_n^*:=\left[\overline{X}_n/\overline{Y}_n\pm t\right]$ cannot have coverage higher than $1-\underline{\alpha}_{n,H}$ since it may be undefined with a probability higher than $\underline{\alpha}_{n,H}$.

□

5.10.5.1 Proof of Lemma 5.9

We consider the following distribution on $\mathbb{R}$

$$P_{n,l_{Y,n},c,\xi}:=\left(\frac{c}{n}\right)^{1/n}\delta_{\{0\}}+\frac{1}{2}\left(1-\left(\frac{c}{n}\right)^{1/n}\right)\delta_{\{y_{c-}\}}+\frac{1}{2}\left(1-\left(\frac{c}{n}\right)^{1/n}\right)\delta_{\{y_{c+}\}},$$

where $c\in(0,n)$ is some constant to be chosen later, $y_{c-}:=l_{Y,n}(1-\xi)/(1-(c/n)^{1/n})$ and $y_{c+}:=l_{Y,n}(1+\xi)/(1-(c/n)^{1/n})$. Let $Y_{1,n}\sim P_{n,l_{Y,n},c,\xi_n}$. Observe that $\mathbb{E}\left[Y_{1,n}\right]=l_{Y,n}$. With the choice

$$c=c_n:=n\left(1-\frac{l_{Y,n}}{b_{Y,n}-a_{Y,n}}\left(1+\xi\right)\right)^n,$$

we have $y_{c+} = b_{Y,n} - a_{Y,n}$. Note that $C_{n,\alpha}$ is strictly positive, because $1 - \frac{l_{Y,n}}{b_{Y,n}-a_{Y,n}}(1+\xi_n) > 0$. This is equivalent to $b_{Y,n} - a_{Y,n}/l_{Y,n} > 1 + \xi_n$, which is true by assumption.

Consider now the following product measure on $\mathbb{R}^2$ defined by $P_n := \left(0.5\delta_{\{0\}} + 0.5\delta_{\{b_{X,n}-a_{X,n}\}}\right) \otimes P_{n,l_{Y,n},c_n,\xi}$. Let $(X_{i,n}, Y_{i,n})_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_n$. These random vectors satisfy $\mathbb{E}[Y_{1,n}] = l_{Y,n}$, $(\max - \min)[Y_{1,n}] = b_{Y,n} - a_{Y,n}$ and $(\max - \min)[X_{1,n}] = b_{X,n} - a_{X,n}$. The next step is to build a lower bound on the event $\{\overline{Y}_n = 0\}$.

The assumption that $(X_{i,n}, Y_{i,n})_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_n$ and the construction of $P_{n,l_{Y,n},c_n,\xi}$ imply that

$$\mathbb{P}\left(\overline{Y}_n = 0\right) = \frac{c_n}{n} = \left(1 - \frac{l_{Y,n}}{b_{X,n} - a_{X,n}}(1+\xi)\right)^n = \tilde{\alpha}_{n,H}(\xi).$$

□

5.11 Additional simulations

This section complements the simulations presented in the main body of the article using different distributions for the variables in the numerator and in the denominator.

In this setting of simulations, we use the best bounds by setting the constants $l_{Y,n}$ and $u_{Y,n}$ that define our class of distributions equal to the actual corresponding moments (respectively the expectation for $l_{Y,n}$ and the second moment for $u_{Y,n}$). That is we use $\overline{n}_\alpha = 2\mathbb{V}[Y]/(\alpha\mathbb{E}[Y]^2)$ or $\overline{\alpha}_n = 2\mathbb{V}[Y]/(n\mathbb{E}[Y]^2)$. In practice, our rule-of-thumb uses the plug-in version of those quantities replacing the theoretical unknown moments by their empirical counterparts as explained in Section 5.5.3.

The following Figures are similar to Figures 5.6 and 5.7. They show the $c(n, P)$ of the asymptotic CIs based on the delta method as a function of the sample size n and also reports $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, with α chosen according to the desired nominal level (equal to $1-\alpha$) and $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$. Consequently, the titles of the figures only indicate the specification used for $P_{X,Y,n}$, the nominal pointwise asymptotic level $1-\alpha$, and the number of repetitions used to approximate the probability $c(n, P)$.

With discrete distributions for the variable in the denominator, it may happen that $\overline{Y}_n = 0$, all the more so as the expectation and the sample size are low typically. As discussed at the end of Section 5.2, confidence intervals are said to be undefined when $\overline{Y}_n = 0$. In such cases, for any value $a \in \mathbb{R}$, it is undefined whether a belongs or not to the CIs. Consequently, whenever the sample drawn is such that $\overline{Y}_n = 0$ in the simulations, we count the draw as a no coverage occurrence in the Monte Carlo estimation of $c(n, P)$. In other words, this quantity is approximated as an average over M repetitions and the repetitions for which $\overline{Y}_n = 0$ account for 0 in this average.[20]

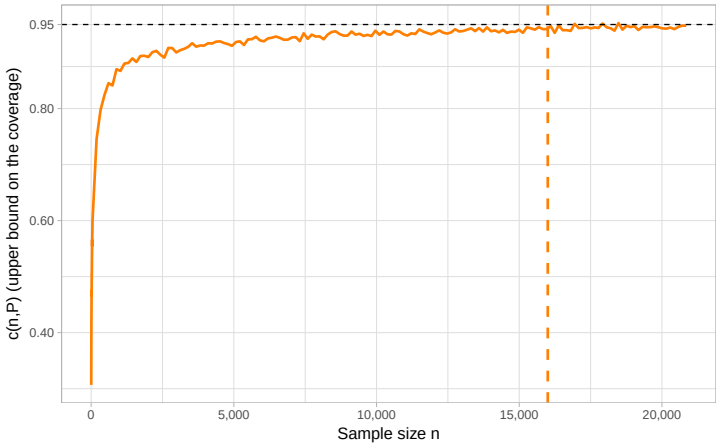

Figure 5.9 – Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}(1,1) \otimes \mathcal{N}(0.05,1)$; $1-\alpha = 0.95$; 5,000 repetitions used.

5.11.1 Gaussian distributions

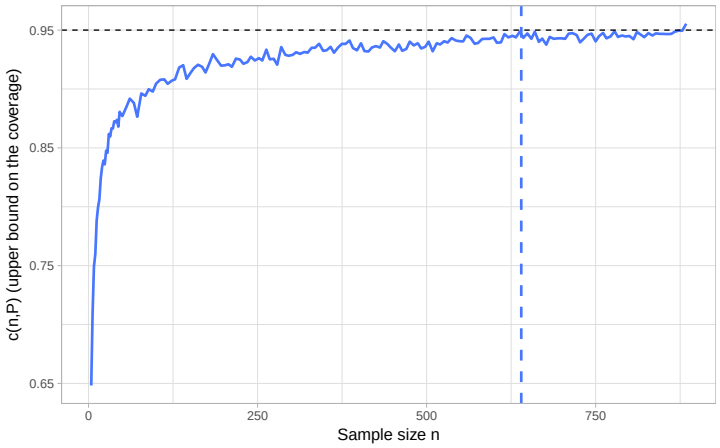

Figure 5.10 – Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}(1,1) \otimes \mathcal{N}(0.25,1)$; $1-\alpha = 0.95$; 5,000 repetitions used.

5.11.2 Student distributions

The specification here is two Student distributions, both in the numerator and in the denominator. Standard Student distributions are centered. We use therefore translated versions by simply adding the expectations in order to avoid a null denominator for the ratio of expectations of interest. Below, $\mathcal{T}(\mu,\nu)$ denotes the distribution of a translated standard Student variable: $\mu + T$ where T is distributed according to a Student distribution with ν degrees of freedom. To satisfy Assumption 5.1, we need finite variance: we use degrees of freedom strictly higher than 2 for this purpose.

5.11.3 Exponential distributions

The specification here is two exponential distributions, both in the numerator and in the denominator. The case of the exponential is specific as a unique parameter determines both the expectation and the variance of the distribution.

[20] Note that in some specifications, a substantial part of the repetitions yield $\overline{Y}_n = 0$. For instance, with Bernoulli distributions, for n smaller than 10 and the expectation at the denominator equal to 0.01, around 10% only of the repetitions display $\overline{Y}_n \neq 0$.

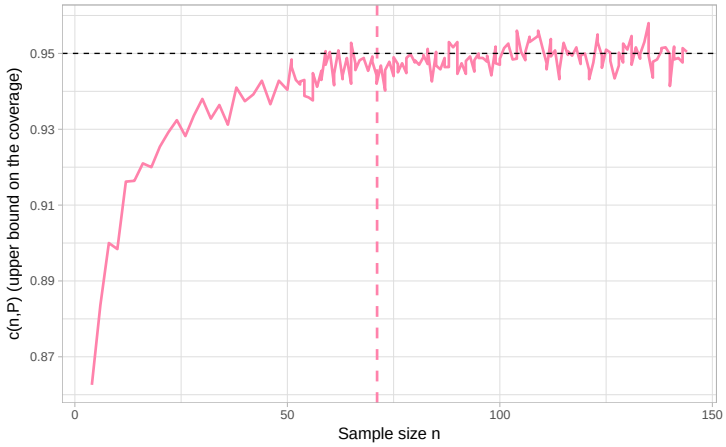

Figure 5.11 – Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}(1,1) \otimes \mathcal{N}(0.75,1)$; $1-\alpha = 0.95$; 5,000 repetitions used.

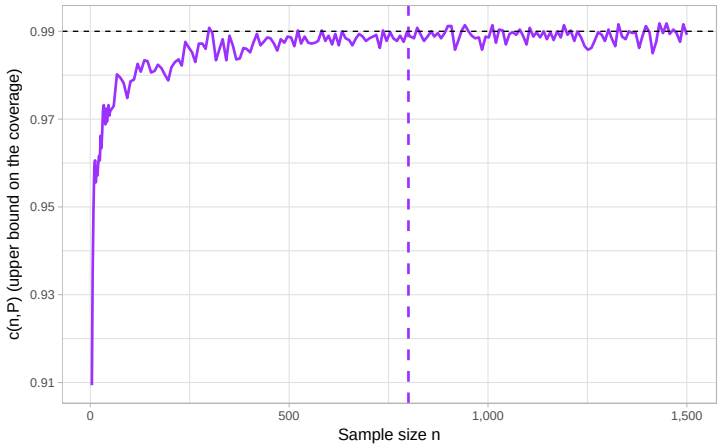

Figure 5.12 – Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}_2$ bivariate Gaussian with $\mathbb{E}[X] = 0.5$, $\mathbb{E}[Y] = 0.5$, $\mathbb{V}[X] = 2$, $\mathbb{V}[Y] = 1$, $\mathbb{C}orr(X,Y) = -0.3$; $1-\alpha = 0.99$; 5,000 repetitions used.

More precisely, the variance is equal to the square of the expectation. Consequently, whatever the parameter of the exponential distribution in the denominator, we have $\overline{n}_\alpha = 4/\alpha$. Previous simulations suggest that the closer the expectation in the denominator to 0, the larger the sample size required for the asymptotic approximation to hold. At first sight, we might thus be worried for the usefulness of our rule-of-thumb to obtain $\overline{n}_\alpha$ independent of $\mathbb{E}[Y]$. Yet, with exponential distributions, the lower the expectation, the lower is the variance too. Intuitively, the lower variance will compensate having an expectation closer to 0. The previous statement that links the closeness to 0 of the expectation in the denominator and the sample size required to reach the asymptotic approximation presupposes keeping fixed the variance. It cannot be anymore for exponential distributions.

The simulations reveal that the convergence of the coverage of the asymptotic confidence intervals toward their nominal level happens for n around one hundred fifty and has the same pattern whatever the expectation of the exponential distribution in the denominator. Our rule-of-thumb $\overline{n}_\alpha$ appears to be a bit small. Nonetheless, it is coherent that it is constant across the value of $\mathbb{E}[Y]$.

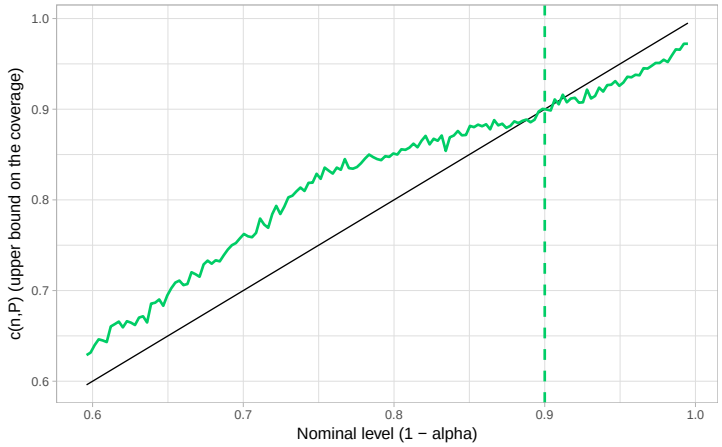

Figure 5.13 – Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}(1,1) \otimes \mathcal{N}(0.1,1)$; $n = 2,000$; 5,000 repetitions used.

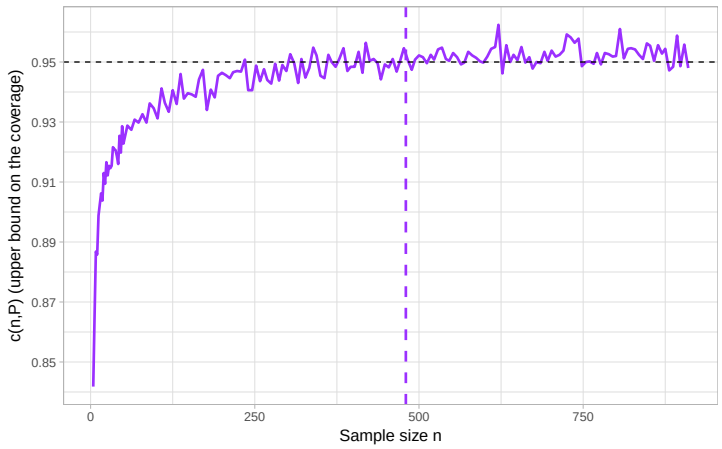

Figure 5.14 – Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{T}(0.5,3) \otimes \mathcal{T}(0.5,3)$; $1-\alpha = 0.95$; 5,000 repetitions used.

5.11.4 Pareto distributions

The specification here is two Pareto distributions, both in the numerator and in the denominator. Pareto distributions have support in $\mathbb{R}_+^*$. They would fall in the easier case when the support of the denominator is well separated from 0. To assess the dependability of our rule-of-thumb in the general case, we use translated Pareto distributions. In what follows, the notation $\mathcal{P}areto(\mathbb{E}[Y], \tau, \gamma)$ denotes the distribution of a random variable that follows a Pareto distribution with shape parameter equal to γ translated such that its support is $(\tau, +\infty)$ and its expectation is $\mathbb{E}[Y]$. A variable that is distributed according to $\mathcal{P}areto(\mathbb{E}[Y], \tau, \gamma)$ is equal in distribution to $P + (\mathbb{E}[Y] - \gamma t_Y)/(\gamma - 1)$ with $t_Y = (\mathbb{E}[Y] - \tau) \times (\gamma - 1)$ and P a usual Pareto distribution with support or scale parameter t_Y and shape parameter γ, that is P has the density $x \mapsto \mathbb{1}\{x \geq t_Y\} \times \gamma t_Y^\gamma / x^{\gamma+1}$ with respect to Lebesgue measure.

5.11.5 Bernoulli distributions

Figure 5.20 is the equivalent of Figure 5.1 with Bernoulli distributions. The following graphs illustrate the use of $\overline{n}_\alpha$ to appraise the reliability of the asymptotic confidence based on the delta method. In practice a plug-in strategy has to be used to compute $\overline{n}_\alpha$ and, in the setting of simulations, we simply use the known moments and bounds of the DGP used in the simulation. With two Bernoulli variables in the numerator and the denominator, we are both in the BC and the “Hoeffding” cases. Thus, we show both the one

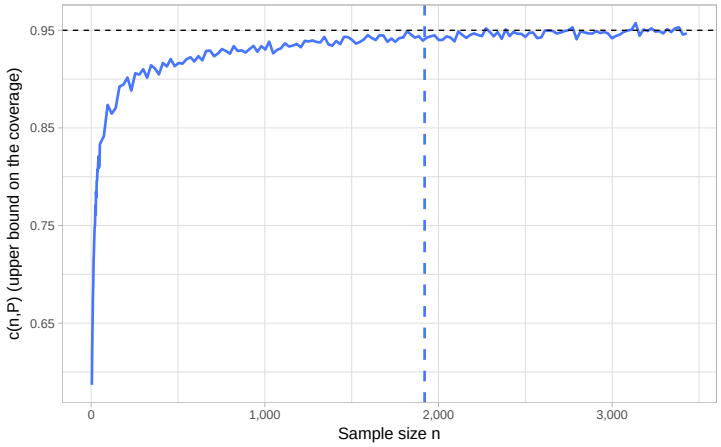

Figure 5.15 – Specification: $\forall n \in \mathbb{N}^*$, the marginal distributions of X and Y are $\mathcal{T}(1,3)$ and $\mathcal{T}(0.25,3)$ respectively and are simulated using a Gaussian copula to have $\mathbb{C}orr(X,Y) \approx 0.5$.; $1-\alpha = 0.95$; 5,000 repetitions used.

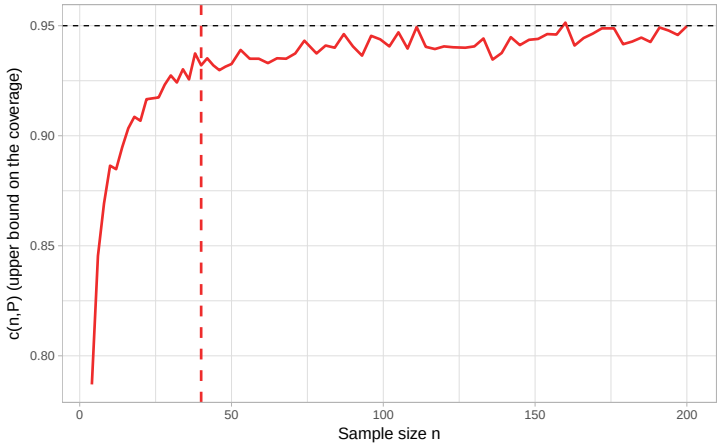

Figure 5.16 – Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{E} \otimes \mathcal{E}$ with $\mathbb{E}[X] = 1$ and $\mathbb{E}[Y] = 0.01$; $1-\alpha = 0.95$; 5,000 repetitions used.

obtained in the BC case $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$ with a dashed vertical line (Theorem 5.4) and the one obtained in the "Hoeffding" case $\overline{n}_{\alpha,H} := \ln(4/\alpha)/\gamma(Y_{1,n})$, setting here $a_{Y,n} = 0$, $b_{Y,n} = 1$ and $l_{Y,n} = \mathbb{E}[Y]$, with a dotted vertical line (Theorem 5.9).

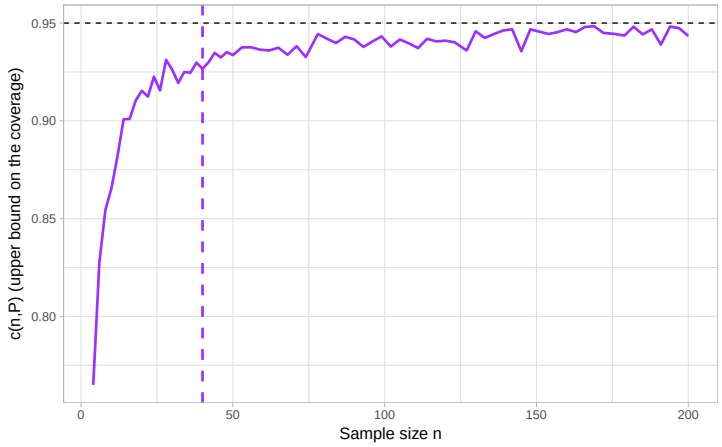

Figure 5.17 – Specification: $\forall n \in \mathbb{N}^*$, the marginal distributions of X and Y are two exponentials with $\mathbb{E}[X] = 1$ and $\mathbb{E}[Y] = 0.5$ and are simulated using a Gaussian copula to have $\mathbb{C}orr(X, Y) \approx 0.75$.; $1 - \alpha = 0.95$; 5,000 repetitions used.

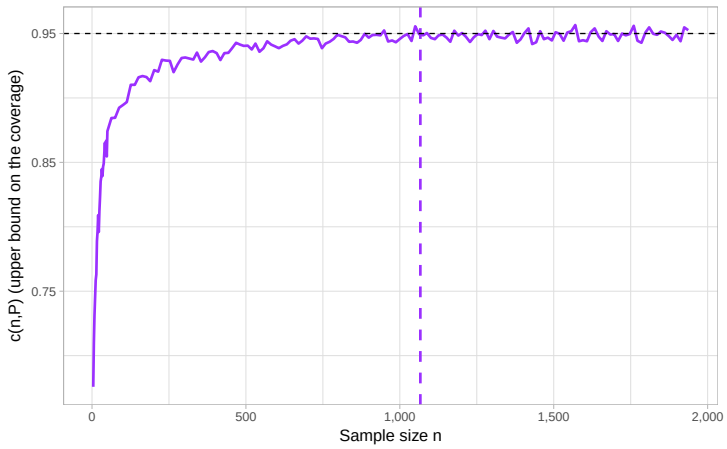

Figure 5.18 – Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{P}areto(1, -1.5, 5) \otimes \mathcal{P}areto(\mathbb{E}[Y], -1.5, 5)$, with $\mathbb{E}[Y] = 0.5$; $1 - \alpha = 0.95$; 5,000 repetitions used.

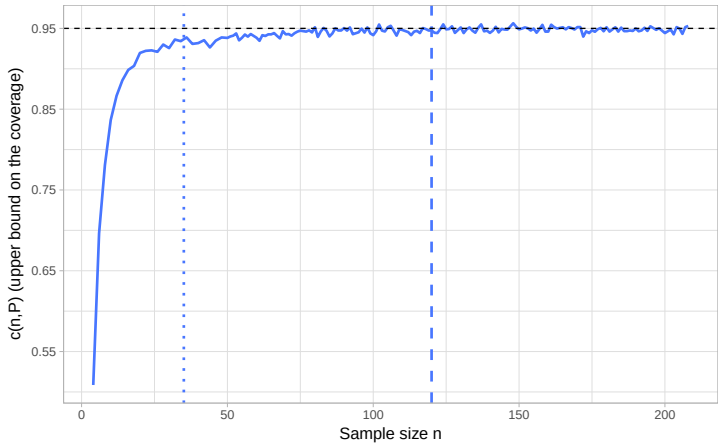

Figure 5.21 – Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{B}(0.5) \otimes \mathcal{B}(0.25)$; $1 - \alpha = 0.95$; 5,000 repetitions used.

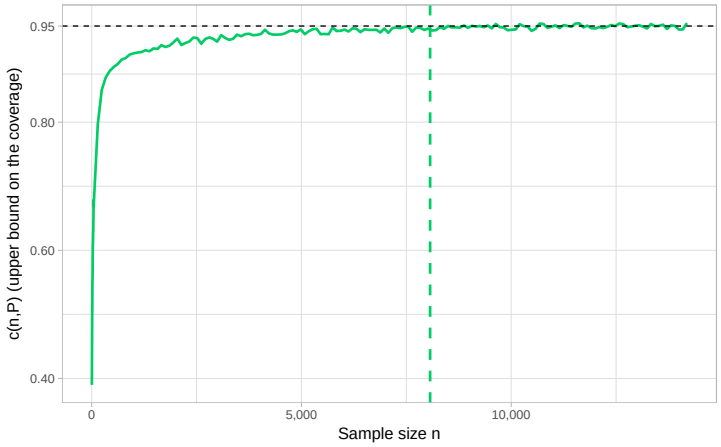

Figure 5.19 – Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{P}areto(1, -1.5, 5) \otimes \mathcal{P}areto(\mathbb{E}[Y], -1.5, 5)$, with $\mathbb{E}[Y] = 0.1$; $1 - \alpha = 0.95$; 5,000 repetitions used.

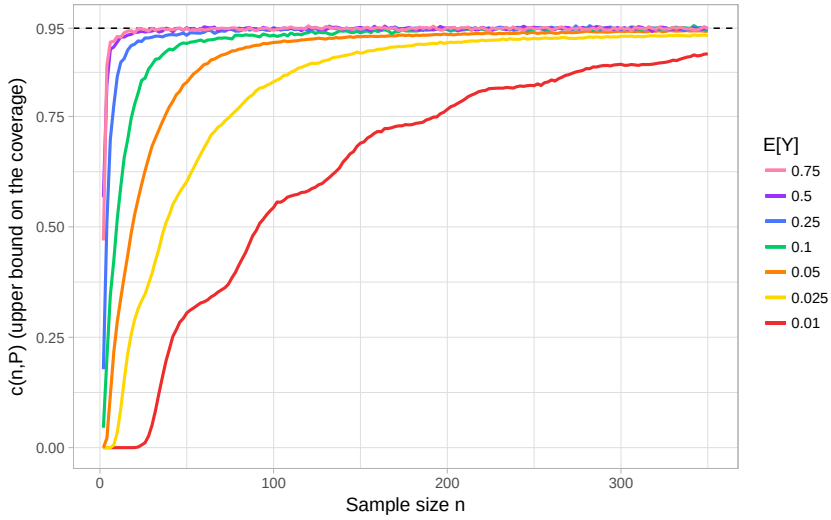

Figure 5.20 – $c(n, P)$ of the CIs based on the delta method as a function of n.
Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{B}(0.5) \otimes \mathcal{B}(\mathbb{E}[Y])$. The nominal pointwise asymptotic level is set to 0.95. 10,000 repetitions used.

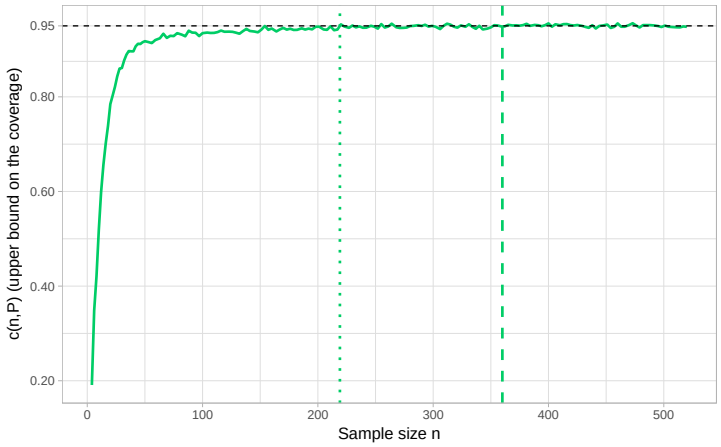

Figure 5.22 – Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{B}(0.5) \otimes \mathcal{B}(0.1)$; $1 - \alpha = 0.95$; 5,000 repetitions used.

5.11.6 Poisson distributions

The specification here considers two variables distributed according to a Poisson distribution, both in the numerator and in the denominator.

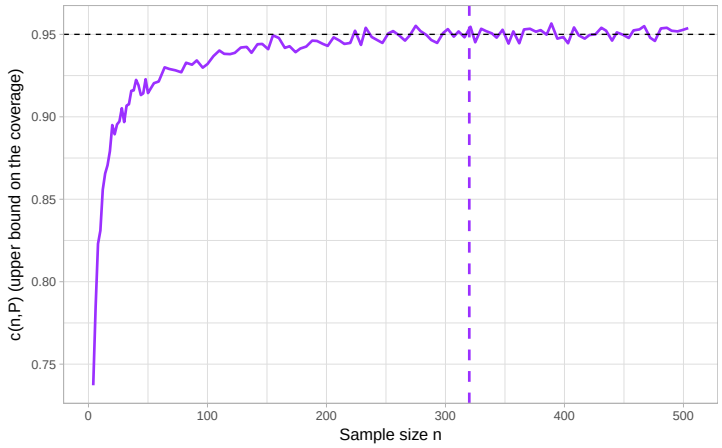

Figure 5.23 – Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{P}oisson(0.5, 2) \otimes \mathcal{P}oisson(0.5, 2)$; $1 - \alpha = 0.95$; 5,000 repetitions used.

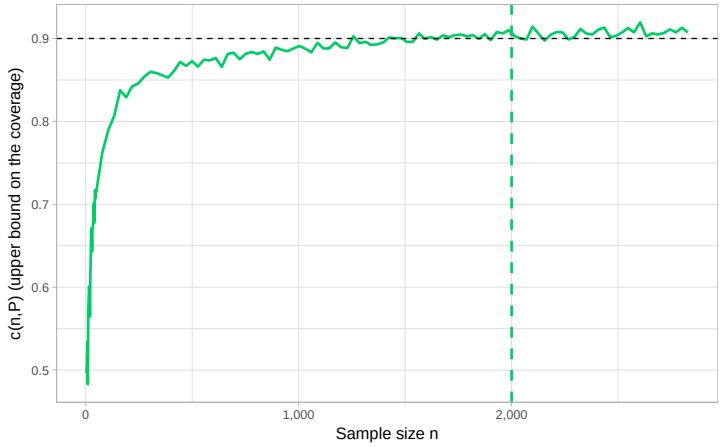

Figure 5.24 – Specification: $\forall n \in \mathbb{N}^*$, the marginal distributions of X and Y are respectively $\mathcal{P}oisson(0.5, 2)$ and $\mathcal{P}oisson(0.1, 1)$ and are simulated using a Gaussian copula to have $\mathbb{C}orr(X, Y) \approx 0.6$; $1 - \alpha = 0.9$; 5,000 repetitions used.

A Poisson distribution is entirely defined by its positive real parameter, which is equal to both its expectation and its variance. Consequently, to have denominator close to 0, we would need small variance too, as in the exponential specification (see Section 5.11.3). In order to disentangle expectation and variance, we use below translated Poisson variables. More precisely, the notation $\mathcal{P}oisson(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+^*$, denotes a distribution alike to a Poisson, with parameter and variance equal to σ^2 but translated such that its expectation is μ. That is a variable distributed according to $\mathcal{P}oisson(\mu, \sigma^2)$ is equal in distribution to $P + (\mu - \sigma^2)$ with P a standard Poisson distribution with parameter σ^2 - that is with density with respect to the counting measure equal to $(\sigma^2)^k \exp(-\sigma^2)/(k!)$ for every $k \in \mathbb{N}$. Thus, a $\mathcal{P}oisson(\mu, \sigma^2)$ has expectation μ and variance σ^2.

5.11.7 Delta method and nonparametric percentile bootstrap confidence intervals

The two following figures are the equivalent to Figure 5.4 with different values of C. They illustrate that the lower C, the lower the signal-to-noise ratio in the denominator, hence the more difficult in some sense is the estimation of θ_n. This is illustrated by the fact that, all other things equal, larger C basically translates $c(n, P)$ upward as revealed by the series of Figures 5.4, 5.25, and 5.26.

These three figures all report the $c(n, P)$ of the CIs based on the delta method (in blue) and of the CIs constructed with Efron's non parametric bootstrap using 2,000 bootstrap replications (in red) with the specification $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}(1, 1) \otimes \mathcal{N}(Cn^{-b}, 1)$, with $b \in \{0, 0.25, 0.5, 0.75\}$. For the three of them, the nominal pointwise asymptotic level is set to 0.95 and for each pair (b, n), the coverage is obtained as the mean over 5,000 repetitions.

Figure 5.25 – delta method in blue; Efron's percentile bootstrap in red; $C = 0.2$.

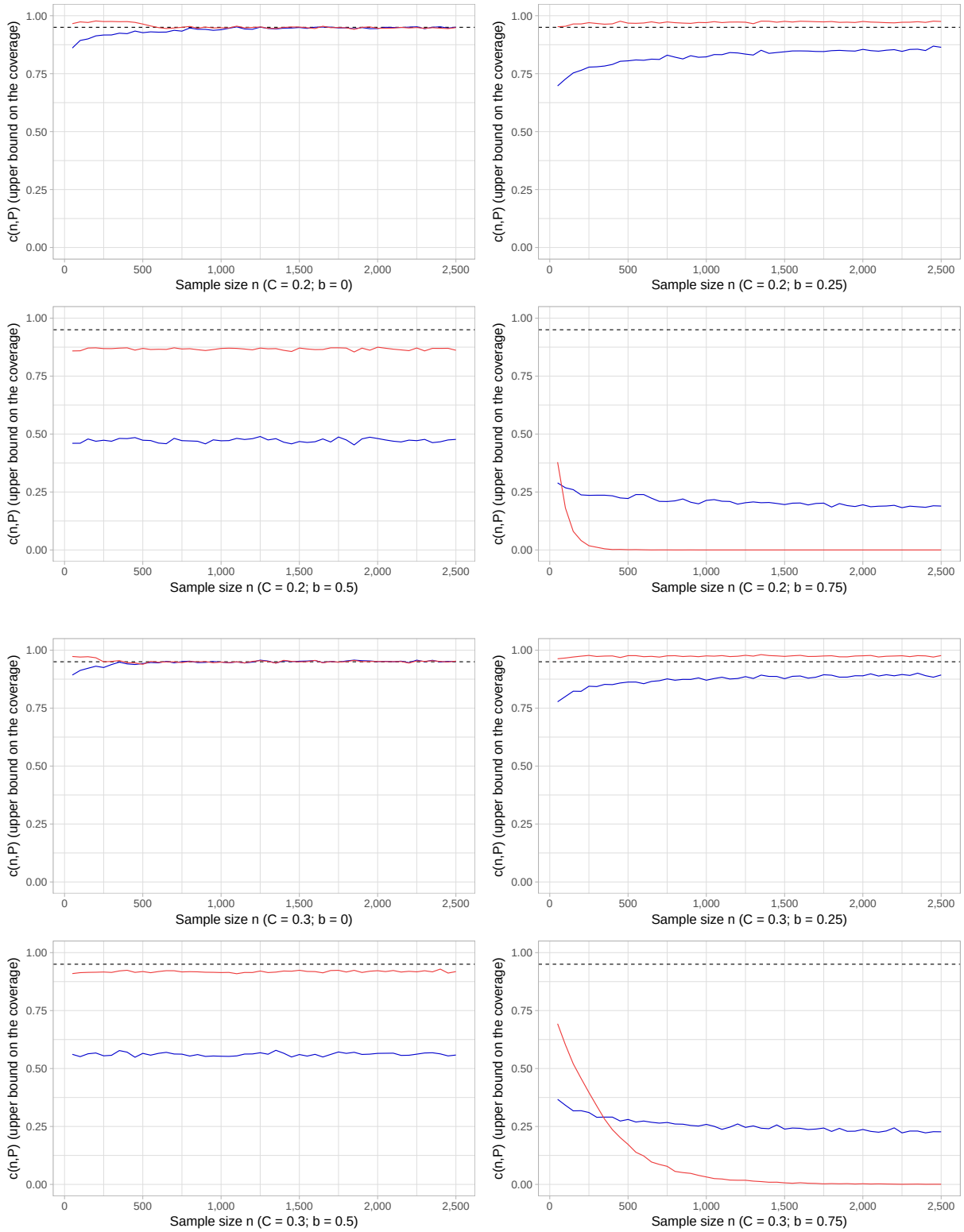

Figure 5.26 – delta method in blue; Efron's percentile bootstrap in red; $C = 0.3$.

Chapter 6

Fuzzy Differences-in-Differences with Stata

Abstract

Differences-in-differences (DID) is a method to evaluate the effect of a treatment. In its basic version, a "control group" is untreated at two dates, whereas a "treatment group" becomes fully treated at the second date. However, in many applications of this method, the treatment rate only increases more in the treatment group. In such fuzzy designs, [53] propose various estimands that identify local average and quantile treatment effects under different assumptions. They also propose estimands that can be used in applications with a non-binary treatment, multiple periods and groups and covariates. This paper presents the Stata command `fuzzydid`, which computes the various corresponding estimators. We illustrate the use of the command by revisiting [76].

Keywords: differences-in-differences, fuzzy designs, local average treatment effects, local quantile treatment effects.
Based on [58] : de Chaisemartin, C., D'Haultfœuille, X. & Guyonvarch Y., Fuzzy Differences-in-Differences with Stata.

6.1 Introduction

Differences-in-differences (DID) is a method to evaluate the effect of a treatment when experimental data are not available. In its basic version, a "control group" is untreated at two dates, whereas a "treatment group" becomes fully treated at the second date. However, in many applications of the DID method the treatment rate increases more in some groups than in others, but there is no group that goes from fully untreated to fully treated, and there is also no group that remains fully untreated. In such fuzzy designs, a popular estimator of treatment effects is the DID of the outcome divided by the DID of the treatment, the so-called Wald-DID.

As shown by [53], the Wald-DID identifies a local average treatment effect (LATE) if two assumptions on treatment effects are satisfied. First, the effect of the treatment should not vary over time. Second, when the treatment increases both in the treatment and in the control group, treatment effects should be equal in these two groups. [53] also propose two alternative estimands of the same LATE. These estimands do not rely on any assumption on treatment effects, and they can be used when the share of

treated units is stable in the control group. The first one, the time-corrected Wald ratio (Wald-TC), relies on common trends assumptions within subgroups of units sharing the same treatment at the first date. The second one, the changes-in-changes Wald ratio (Wald-CIC), generalizes the changes-in-changes estimand introduced by [12] to fuzzy designs. Finally, under the same assumptions as those used for the Wald-CIC, local quantile treatment effects (LQTE) are also identified.

In this paper, we describe the `fuzzydid` Stata command, which computes the estimators corresponding to these estimands and performs inference on the LATE and LQTE using the bootstrap. In the computation of standard errors and confidence intervals, clustering along one dimension can be allowed for. Equality tests between the Wald-DID, Wald-TC, and Wald-CIC and placebo tests can also be performed. This turns out to be important for choosing between these different estimands, as they identify the LATE under different sets of assumptions.

The identification results mentioned above hold with a control group where the share of treated units does not change over time, a binary treatment, no covariates, and two groups and two periods. Nonetheless, they can be extended in several directions. First, under the same assumptions as those underlying the Wald-TC estimand, the LATE of treatment group switchers can be bounded when the share of treated units changes over time in the control group. Second, non-binary treatments can be easily handled by just modifying the parameter of interest. Third, when the assumptions are more credible conditional on some controls, it is possible to modify the Wald-DID, Wald-TC, and Wald-CIC estimands to incorporate such controls. The `fuzzydid` command handles all these extensions.

Finally, results can be extended to applications with multiple periods and groups. Those are very prevalent in applied work, and researchers then estimate treatment effects through linear regressions including time and group fixed effects. [52] show that around 19% of all empirical papers published by the American Economic Review between 2010 and 2012 make use of this research design. This paper also shows that these regressions are extensions of the Wald-DID to multiple periods and groups, and that they identify weighted averages of LATE, with possibly many negative weights.[1] As a result, they do not satisfy the no-sign reversal property: the coefficient of the treatment variable in those regressions may be negative even if the treatment effect is positive for every unit in the population. On the other hand, the Wald-DID, Wald-TC, and Wald-CIC estimands can be extended to applications with multiple groups and periods, and they then identify a LATE under the same assumptions as in the two groups and two periods case. Again, the `fuzzydid` command computes the corresponding estimators.

The remainder of the paper is organized as follows. Section 6.2 presents the estimands and estimators considerered by [53] in the simplest set-up with two groups and periods, a binary treatment and no covariates. Section 6.3 discusses the various extensions covered by the command. Section 6.4 presents the `fuzzydid` Stata command. Section 6.5 illustrates the command by revisiting [76], who estimate the effect of newspapers on electoral participation. Section 6.6 presents the finite sample performances of the various estimators through Monte Carlo simulations. Section 6.7 concludes.

[1] A Stata command computing these weights is available on the authors' webpages.

6.2 Set-up

6.2.1 Parameters of interest, assumptions, and estimands

We seek to identify the effect of a treatment D on some outcome. In this section, we assume that D is binary.[2] $Y(1)$ and $Y(0)$ denote the two potential outcomes of the same individual with and without treatment, while $Y = Y(D)$ denotes the observed outcome. We assume the data can be divided into time periods represented by a random variable $T \in \{0, ..., \bar{t}\}$, and into groups represented by a random variable $G \in \{0, ..., \bar{g}\}$. We start by considering the simple case where $\bar{t} = \bar{g} = 1$, thus implying that there are two groups and two periods. In such a case, $G = 1$ (resp. $G = 0$) for units in the treatment (resp. control) group.

We use the following notation hereafter. For any random variable R, Supp(R) denotes its support. R_{gt} and R_{dgt} are two other random variables such that $R_{gt} \sim R|\, G = g, T = t$ and $R_{dgt} \sim R|\, D = d, G = g, T = t$, where $\sim$ denotes equality in distribution. For any event or random variable A, F_R and $F_{R|A}$ denote respectively the cumulative distribution function (cdf) of R and its cdf conditional on A. Finally, for any increasing function F on the real line, we let $F^{-1}(q) = \inf\{x \in \mathbb{R} : F(x) \geq q\}$. In particular, F_R^{-1} is the quantile function of R.

We maintain Assumptions 6.1-6.3 below in most of the paper.

Assumption 6.1. *(Fuzzy design)*
$E(D_{11}) > E(D_{10})$*, and* $E(D_{11}) - E(D_{10}) > E(D_{01}) - E(D_{00})$*.*

Assumption 6.2. *(Stable percentage of treated units in the control group)*
For all $d \in$ *Supp*(D)*,* $P(D_{01} = d) = P(D_{00} = d) \in (0, 1)$*.*

Assumption 6.3. *(Treatment participation equation)*
There exist $D(0), ..., D(\bar{t})$ *such that* $D = D(T)$*,* $D(t) \perp\!\!\!\perp T|G$ *(*$t \in \{0, ..., \bar{t}\}$*) and for all* $t \in \{1, ..., \bar{t}\}$*,*

$$P(D(t) \geq D(t-1)|G) = 1 \;\; \textit{or}\; P(D(t) \leq D(t-1)|G) = 1.$$

In standard "sharp" designs, we have $D = G \times T$, meaning that only observations in the treatment group and in period 1 get treated. With Assumption 6.1, we consider instead "fuzzy" settings where $D \neq G \times T$ in general, but where the treatment group experiences a higher increase of its treatment rate between period 0 and 1. Assumption 6.2 requires that the treatment rate remain constant in the control group, and be strictly included between 0 and 1. This assumption is testable. Assumption 6.3 is equivalent to the latent index model $D = \mathbb{1}\{V \geq v_{GT}\}$ (with $V \perp\!\!\!\perp T|G$) considered in [53]. In repeated cross sections, $D(t)$ denotes the treatment status of a unit at period t, and only $D = D(T)$ is observed. In single cross sections where cohort of birth plays the role of time, $D(t)$ denotes instead the potential treatment of a unit had she been born at $T = t$. Here again, only $D = D(T)$ is observed.

We consider the subpopulation $S = \{D(0) < D(1), G = 1\}$, called hereafter the treatment group switchers. Our parameters of interest are their Local Average Treatment Effect (LATE) and Local Quantile Treatment Effects (LQTE), which are respectively defined by

$$\begin{aligned} \Delta &= E\left(Y(1) - Y(0)|S, T = 1\right), \\ \tau_q &= F^{-1}_{Y(1)|S,T=1}(q) - F^{-1}_{Y(0)|S,T=1}(q), \quad q \in (0, 1). \end{aligned}$$

[2] We still define our assumptions and estimands for any scalar treatment, to avoid redefining them when we will extend our results to non-binary treatments.

We now introduce the main estimands considered in [53]. We start by considering the three estimands of Δ. The first is the Wald-DID defined by

$$W_{DID} = \frac{E(Y_{11}) - E(Y_{10}) - (E(Y_{01}) - E(Y_{00}))}{E(D_{11}) - E(D_{10}) - (E(D_{01}) - E(D_{00}))}.$$

W_{DID} is the coefficient of D in a 2SLS regression of Y on D with G and T as included instruments, and $G \times T$ as the excluded instrument.

The second estimand of Δ is the time-corrected Wald ratio (Wald-TC) defined by

$$W_{TC} = \frac{E(Y_{11}) - E(Y_{10} + \delta_{D_{10}})}{E(D_{11}) - E(D_{10})},$$

where $\delta_d = E(Y_{d01}) - E(Y_{d00})$, for $d \in \text{Supp}(D)$. Without the $\delta_{D_{10}}$ term, W_{TC} would correspond to the coefficient of D in a 2SLS regression of Y on D using T as the excluded instrument, within the treatment group. δ_0 (resp. δ_1) measures the evolution of the outcome among untreated (resp. treated) units in the control group. Under the assumption that these evolutions are the same in the two groups (see Assumption 6.4 ' below), the $\delta_{D_{10}}$ term accounts for the effect of time on the outcome in the treatment group.

The third estimand of Δ is the change-in-change Wald ratio (Wald-CIC) defined by

$$W_{CIC} = \frac{E(Y_{11}) - E(Q_{D_{10}}(Y_{10}))}{E(D_{11}) - E(D_{10})},$$

where $Q_d(y) = F^{-1}_{Y_{d01}} \circ F_{Y_{d00}}(y)$ is the quantile-quantile transform of Y from period 0 to 1 in the control group conditional on $D = d$. W_{CIC} is similar to W_{TC}, except that it accounts for the effect of time on the outcome through the quantile-quantile transform instead of the additive term $\delta_{D_{10}}$.

Finally, we consider an estimand of τ_q. Let

$$F_{CIC,d} = \frac{P(D_{11} = d)F_{Y_{d11}} - P(D_{10} = d)F_{Q_d(Y_{d10})}}{P(D_{11} = d) - P(D_{10} = d)}$$

and

$$\tau_{CIC,q} = F^{-1}_{CIC,1}(q) - F^{-1}_{CIC,0}(q).$$

The estimands above identify Δ or τ_q under combinations of the following assumptions.

Assumption 6.4. *(Common trends)*

For all $t \in \{1, ..., \bar{t}\}$, $E(Y(0)|G, T = t) - E(Y(0)|G, T = t - 1)$ *does not depend on* G.

Assumption 6.4'. *(Conditional common trends)*

For all $d \in$ *Supp*(D) *and all* $t \in \{1, ..., \bar{t}\}$, $E(Y(d)|G, T = t, D(t-1) = d) - E(Y(d)|G, T = t-1, D(t-1) = d)$ *does not depend on* G.

Assumption 6.5. *(Stable treatment effect over time)*

For all $d \in$ *Supp*(D) *and all* $t \in \{1, ..., \bar{t}\}$, $E(Y(d) - Y(0)|G, T = t, D(t-1) = d) = E(Y(d) - Y(0)|G, T = t - 1, D(t-1) = d)$.

Assumption 6.6. *(Monotonicity and time invariance of unobservables)*

$Y(d) = h_d(U_d, T)$, *with* $U_d \in \mathbb{R}$ *and* $h_d(u, t)$ *strictly increasing in* u *for all* $(d, t) \in$ *Supp*$(D) \times$ *Supp*(T). *Moreover,* $U_d \perp\!\!\!\perp T|G, D(0)$.

Assumption 6.7. *(Data restrictions)*

1. *Supp*$(Y_{dgt}) =$ *Supp*$(Y) = [\underline{y}, \overline{y}]$ *with* $-\infty \leq \underline{y} < \overline{y} \leq +\infty$, *for* $(d, g, t) \in$ *Supp*$((D, G, T))$.

2. $F_{Y_{dgt}}$ *is continuous on* $\mathbb{R}$ *and strictly increasing on Supp*(Y)*, for* $(d, g, t) \in$ *Supp*$((D, G, T))$*.*

Assumption 6.4 is the usual common trends condition, under which the DID estimand identifies the average treatment effect on the treated in sharp designs where $D = G \times T$. Assumption 6.4 ' is a conditional version of this common trend condition, which requires that the mean of $Y(0)$ (resp. $Y(1)$) among untreated (resp. treated) units at period 0 follow the same evolution in both groups. Assumption 6.5 requires that in each group, the average treatment effect among units treated in period 0 remain stable between periods 0 and 1. Assumption 6.6 requires that potential outcomes be strictly increasing functions of a scalar and stationary unobserved term, as in [12]. Assumption 6.7 is a testable restriction on the distribution of Y that is necessary only for the Wald-CIC and $\tau_{q,CIC}$ estimands.

Theorem 6.1. *[53] Suppose that Assumptions 6.1-6.3 hold.*

1. *If Assumptions 6.4 and 6.5 also hold, then* $W_{DID} = \Delta$*.*

2. *If Assumptions 6.4 ' also hold, then* $W_{TC} = \Delta$*.*

3. *If Assumptions 6.6-6.7 also hold, then* $W_{CIC} = \Delta$ *and* $\tau_{q,CIC} = \tau_q$*.*

Theorem 6.1 gives several sets of conditions under which we can identify Δ, using one of the three estimands above. It also shows that τ_q can be identified under the same conditions as those under which the Wald-CIC identifies Δ. Compared to the Wald-DID, the Wald-TC and Wald-CIC do not rely on the stable treatment effect assumption, which may be implausible. The choice between the Wald-TC and the Wald-CIC estimands should be based on the suitability of Assumption 6.4 ' and 6.6 in the application under consideration. Assumption 6.4 ' is not invariant to the scaling of the outcome, but it only restricts its mean. Assumption 6.6 is invariant to the scaling of the outcome, but it restricts its entire distribution. When the treatment and control groups have different outcome distributions conditional on D in the first period, the scaling of the outcome might have a large effect on the Wald-TC. The Wald-CIC is much less sensitive to the scaling of the outcome, so using this estimand might be preferable. On the other hand, when the two groups have similar outcome distributions conditional on D in the first period, using the Wald-TC might be preferable.

To test the assumptions underlying those estimands, one can test whether they are equal. If they are not, at least one of those assumptions must be violated. An alternative approach is to perform placebo tests. For instance, if three time periods are available ($T =$-1, 0, or 1), and if the treatment rate remains stable in both groups between $T =$-1 and 0, the numerators of the Wald-DID, Wald-TC, and Wald-CIC estimands for those two periods should be equal to zero.

6.2.2 Estimators

We now turn to the estimation of Δ and $\tau_{q,CIC}$ using plug-in estimators of the estimands above. Let $(Y_i, D_i, G_i, T_i)_{i=1...n}$ denote an i.i.d. sample of (Y, D, G, T) and define $\mathcal{I}_{gt} = \{i : G_i = g, T_i = t\}$ and $\mathcal{I}_{dgt} = \{i : D_i = d, G_i = g, T_i = t\}$. Let n_{gt} and n_{dgt} denote the size of $\mathcal{I}_{gt}$ and $\mathcal{I}_{dgt}$, for all $(d, g, t) \in \mathcal{S}(D) \times \{0, 1\}^2$.

First, let

$$\widehat{W}_{DID} = \frac{\frac{1}{n_{11}} \sum_{i \in \mathcal{I}_{11}} Y_i - \frac{1}{n_{10}} \sum_{i \in \mathcal{I}_{10}} Y_i - \frac{1}{n_{01}} \sum_{i \in \mathcal{I}_{01}} Y_i + \frac{1}{n_{00}} \sum_{i \in \mathcal{I}_{00}} Y_i}{\frac{1}{n_{11}} \sum_{i \in \mathcal{I}_{11}} D_i - \frac{1}{n_{10}} \sum_{i \in \mathcal{I}_{10}} D_i - \frac{1}{n_{01}} \sum_{i \in \mathcal{I}_{01}} D_i + \frac{1}{n_{00}} \sum_{i \in \mathcal{I}_{00}} D_i}$$

be the estimator of the Wald-DID. Second, for any $d \in$ Supp(D) let $\widehat{\delta}_d = (1/n_{d01}) \sum_{i \in \mathcal{I}_{d01}} Y_i -$

$(1/n_{d00})\sum_{i\in\mathcal{I}_{d00}} Y_i$. Then, let

$$\widehat{W}_{TC} = \frac{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}} Y_i - \frac{1}{n_{10}}\sum_{i\in\mathcal{I}_{10}}\left[Y_i + \widehat{\delta}_{D_i}\right]}{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}} D_i - \frac{1}{n_{10}}\sum_{i\in\mathcal{I}_{10}} D_i}$$

be the estimator of the Wald-TC. Third, for all $(d,g,t)\in\mathcal{S}(D)\times\{0,1\}^2$, let $\widehat{F}_{Y_{dgt}}(y) = \frac{1}{n_{dgt}}\sum_{i\in\mathcal{I}_{dgt}} \mathbb{1}\{Y_i \leq y\}$ denote the empirical cdf of Y_{dgt}. Let

$$\widehat{Q}_d(y) = \max\left(\widehat{F}^{-1}_{Y_{d01}} \circ \widehat{F}_{Y_{d00}}(y), \min\{Y_i : i\in\mathcal{I}_{d01}\}\right)$$

be the estimator of the quantile-quantile transform Q_d, and let

$$\widehat{W}_{CIC} = \frac{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}} Y_i - \frac{1}{n_{10}}\sum_{i\in\mathcal{I}_{10}}\widehat{Q}_{D_i}(Y_i)}{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}} D_i - \frac{1}{n_{10}}\sum_{i\in\mathcal{I}_{10}} D_i}$$

be the estimator of the Wald-CIC. Finally, let $\widehat{P}(D_{gt}=d) = n_{dgt}/n_{gt}$ and

$$\widehat{F}^{\text{pi}}_{CIC,d} = \frac{\widehat{P}(D_{11}=d)\widehat{F}_{Y_{d11}} - \widehat{P}(D_{10}=d)\widehat{F}_{\widehat{Q}_d(Y_{d10})}}{\widehat{P}(D_{11}=d) - \widehat{P}(D_{10}=d)}.$$

The function $\widehat{F}^{\text{pi}}_{CIC,d}$ is the plug-in estimator of $F_{CIC,d}$ but it has the drawback of not being necessarily a proper cdf. It may not be nondecreasing and may not belong to $[0,1]$. To avoid these issues, we consider a rearranged version $\widehat{F}^{\text{arr}}_{CIC,d}$ of $\widehat{F}^{\text{pi}}_{CIC,d}$, following [42]. Moreover, we let

$$\widehat{F}_{CIC,d}(y) = \max\left(\min(\widehat{F}^{\text{arr}}_{CIC,d}(y), 1), 0\right).$$

With this proper cdf at hand, let

$$\widehat{\tau}_q = \widehat{F}^{-1}_{CIC,d}(q) - \widehat{F}^{-1}_{CIC,d}(q)$$

be the estimator of τ_q.

[53] show that $\widehat{W}_{DID}$, $\widehat{W}_{TC}$, $\widehat{W}_{CIC}$, and $\widehat{\tau}_q$ are root-n consistent and asymptotically normal under standard regularity conditions.[3] [53] also establish the validity of the bootstrap to draw inference on Δ and τ_q based on these estimators. The `fuzzydid` command uses the bootstrap to compute the standard errors of all estimators, and the percentile bootstrap to compute confidence intervals.

6.3 Extensions

6.3.1 Including covariates

The basic set-up can be extended to include covariates. Let X denote a vector of covariates, and for any random variable R, let $m^R_{gt}(x) = E(R_{gt}|X=x)$. Let also $\delta_d(x) = E(Y_{d01}|X=x) - E(Y_{d00}|X=x)$ and $\widetilde{\delta}(x) = E\left(\delta_{D_{10}}(X_{10})|X=x\right]$. Then define

$$W^X_{DID} = \frac{E(Y_{11}) - E(m^Y_{10}(X_{11})) - \left(E(m^Y_{01}(X_{11})) - E(m^Y_{00}(X_{11}))\right)}{E(D_{11}) - E(m^D_{10}(X_{11})) - \left(E(m^D_{01}(X_{11})) - E(m^D_{00}(X_{11}))\right)},$$

$$W^X_{TC} = \frac{E(Y_{11}) - E\left[m^Y_{10}(X_{11}) + \widetilde{\delta}(X_{11})\right]}{E(D_{11}) - E(m^D_{10}(X_{11}))}.$$

[3] [53] consider an estimator of τ_q based on $\widehat{F}^{\text{pi}}_{CIC,d}$ rather than $\widehat{F}_{CIC,d}$. However, these two estimators are equal on any compact set with probability tending to one whenever $F_{CIC,d}$ is strictly increasing. Thus, the two estimators of τ_q also coincide with probability tending to one, and their result also applies to the estimator considered here.

[54] show that W^X_{DID} (resp. W^X_{TC}) identifies Δ under the common support condition Supp$(X_{gt}) =$ Supp(X) for all (g,t) (resp. Supp$(X_{dgt}) =$ Supp(X) for all (d,g,t)) and conditional versions of Assumptions 6.1-6.3 and 6.4-6.5 (resp. 6.4 ').[4]

Let us turn to estimators of W^X_{DID} and W^X_{TC}. We first consider non-parametric estimators. Let us assume that $X \in \mathbb{R}^r$ is a vector of continuous covariates. Adding discrete covariates is easy by reasoning conditional on each corresponding cell. We take an approach similar to, e.g., [74] by estimating in a first step conditional expectations by series estimators. For any positive integer K, let $p^K(x) = (p_{1K}(x), ..., p_{KK}(x))'$ be a vector of basis functions and $P^K_{gt} = (p^K(X_1), ..., p^K(X_n))$. For any random variable R, we estimate $m^R(x) = E(R|X = x)$ by the series estimator

$$\widehat{m}^R(x) = p^{K_n}(x)' \left(P^{K_n} P^{K_n\prime}\right)^- P^{K_n} (R_1, ..., R_n)',$$

where $(.)^-$ denotes the generalized inverse and K_n is an integer. We then estimate $m^R_{gt}(x) = E(R_{gt}|X = x)$ by the series estimator above on the subsample $\{i : G_i = g, T_i = t\}$. $m^R_{dgt}(x) = E(R_{dgt}|X = x)$ is estimated similarly. Then our non-parametric estimators of W^X_{DID} and W^X_{TC} are defined as

$$\widehat{W}^X_{DID,NP} = \frac{\frac{1}{n_{11}} \sum_{i\in\mathcal{I}_{11}} \left[Y_i - \widehat{m}^Y_{10}(X_i) - \widehat{m}^Y_{01}(X_i) + \widehat{m}^Y_{00}(X_i)\right]}{\frac{1}{n_{11}} \sum_{i\in\mathcal{I}_{11}} \left[D_i - \widehat{m}^D_{10}(X_i) - \widehat{m}^D_{01}(X_i) + \widehat{m}^D_{00}(X_i)\right]},$$
$$\widehat{W}^X_{TC,NP} = \frac{\frac{1}{n_{11}} \sum_{i\in\mathcal{I}_{11}} \left[Y_i - \widehat{m}^Y_{10}(X_i) - \widehat{m}^D_{10}(X_i)\widehat{\delta}_1(X_i) - (1 - \widehat{m}^D_{10}(X_i))\widehat{\delta}_0(X_i)\right]}{\frac{1}{n_{11}} \sum_{i\in\mathcal{I}_{11}} \left[D_i - \widehat{m}^D_{10}(X_i)\right]},$$

where $\widehat{\delta}_d(x) = \widehat{m}^Y_{d01}(x) - \widehat{m}^Y_{d00}(x)$. Under regularity conditions, these estimators are root-n consistent and asymptotically normal [see 54, Section 2.3].

Second, we consider semi-parametric estimators of W^X_{DID} and W^X_{TC}. Assume for instance that for $(d,g,t) \in \{0,1\}^3$, $E(Y_{gt}|X) = X'\beta^Y_{gt}$, $E(Y_{dgt}|X) = X'\beta^Y_{dgt}$, and $E(D_{gt}|X) = X'\beta^D_{gt}$. Under this assumption, we have

$$W^X_{DID} = \frac{E(Y_{11}) - E\left(X'_{11}\beta^Y_{10}\right) - \left(E\left(X'_{11}\beta^Y_{01}\right) - E\left(X'_{11}\beta^Y_{00}\right)\right)}{E(D_{11}) - E\left(X'_{11}\beta^D_{10}\right) - \left(E\left(X'_{11}\beta^D_{01}\right) - E\left(X'_{11}\beta^D_{00}\right)\right)}$$
$$W^X_{TC} = \frac{E(Y_{11}) - E\left[X'_{11}\left(\beta^Y_{10} + X'_{11}\beta^D_{10}\left(\beta^Y_{101} - \beta^Y_{100}\right) + (1 - X'_{11}\beta^D_{10})\left(\beta^Y_{001} - \beta^Y_{000}\right)\right)\right]}{E(D_{11}) - E(X'_{11}\beta^D_{10})}.$$

Then, semi-parametric estimators of W^X_{DID} and W^X_{TC} can be defined as

$$\widehat{W}^X_{DID,OLS} = \frac{\sum_{i\in\mathcal{I}_{11}} \left[Y_i - X'_i\widehat{\beta}^Y_{10} - X'_i\widehat{\beta}^Y_{01} + X'_i\widehat{\beta}^Y_{00}\right]}{\sum_{i\in\mathcal{I}_{11}} \left[D_i - X'_i\widehat{\beta}^D_{10} - X'_i\widehat{\beta}^D_{01} + X'_i\widehat{\beta}^D_{00}\right]},$$
$$\widehat{W}^X_{TC,OLS} = \frac{\sum_{i\in\mathcal{I}_{11}} Y_i - \left[X'_i\widehat{\beta}^Y_{10} + X'_i\left(X'_i\widehat{\beta}^D_{10}(\widehat{\beta}^Y_{101} - \widehat{\beta}^Y_{100}) + (1 - X'_i\widehat{\beta}^D_{10})(\widehat{\beta}^Y_{001} - \widehat{\beta}^Y_{000})\right)\right]}{\sum_{i\in\mathcal{I}_{11}} \left[D_i - X'_i\widehat{\beta}^D_{10}\right]},$$

where for $(d,g,t) \in \{0,1\}^3$, $\widehat{\beta}^Y_{gt}$ (resp. $\widehat{\beta}^Y_{dgt}$) denotes the coefficient of X in an OLS regression of Y on X in the subsample $\mathcal{I}_{gt}$ (resp. $\mathcal{I}_{dgt}$), and $\widehat{\beta}^D_{gt}$ denotes the coefficient of X in an OLS regression of D on X in the subsample $\mathcal{I}_{gt}$. When either Y or D is binary, one might prefer to posit a probit or a logit model for its conditional expectation functions in the various subsamples. Other semi-parametric estimators can be defined accordingly.

Finally, researchers may sometimes wish to include a large set of controls in their estimation, which may lead to violations of the common support assumptions Supp$(X_{gt}) =$ Supp(X) and Supp$(X_{dgt}) =$

[4] [54] also propose a Wald-CIC estimand with covariates, but the corresponding estimator is not computed by the `fuzzydid` command.

Supp(X).[5] For instance, when the researcher wants to estimate the Wald-DID, there might be values of X for which all units belong to the treatment group, thus implying that for those values there are no control units to which the trends experienced by treatment group units can be compared. Let x_0 denote one such problematic value, i.e. $x_0 \in$ Supp(X_{11}) but $E(Y_{0t}|X = x_0)$ and $E(D_{0t}|X = x_0)$ are not defined for some $t \in \{0, 1\}$. To avoid dropping treatment group units with $X = x_0$, we use all control units to predict their counterfactual trends. Namely, in W^X_{DID} we replace $E(Y_{01}|X = x_0) - E(Y_{00}|X = x_0)$ and $E(D_{01}|X = x_0) - E(D_{00}|X = x_0)$ by $E(Y_{01}) - E(Y_{00})$ and $E(D_{01}) - E(D_{00})$. If instead, the researcher wants to estimate the Wald-TC, the same principle applies.

6.3.2 Multiple periods and groups

We now extend our initial setting to multiple periods and groups. We first define, at each period $t \in \{1, ..., \bar{t}\}$, the following "supergroup" variable

$$G^*_t = 1\{E(D_{gt}) > E(D_{gt-1})\} - 1\{E(D_{gt}) < E(D_{gt-1})\}.$$

Let $\mathcal{T} = \left\{t \in \{1, ..., \bar{t}\} : P(G^*_t = 0) > 0\right\}$ denote the subset of periods t for which there exists at least one group with stable treatment rate between $t-1$ and t. We let $S = \{D(T) \neq D(T-1), T \in \mathcal{T}\}$ denote the population of units switching between $T-1$ and $T \in \mathcal{T}$ and define Δ in this set-up as $\Delta = E[Y(1) - Y(0)|S]$. For any random variable R and any $(d, g, t) \in \{0, 1\} \times \{-1, 1\} \times \mathcal{T}$, we also define the following quantities:

$$\begin{aligned}
DID^*_R(g,t) =& E(R|G^*_t = g, T = t) - E(R|G^*_t = g, T = t-1) \\
&- (E(R|G^*_t = 0, T = t) - E(R|G^*_t = 0, T = t-1)), \\
\delta^*_{dt} =& E(Y|D = d, G^*_t = 0, T = t) - E(Y|D = d, G^*_t = 0, T = t-1), \\
Q^*_{dt}(y) =& F^{-1}_{Y|D=d,G^*_t=0,T=t} \circ F_{Y|D=d,G^*_t=0,T=t-1}(y), \\
W^*_{DID}(g,t) =& \frac{DID^*_Y(g,t)}{DID^*_D(g,t)}, \\
W^*_{TC}(g,t) =& \frac{E(Y|G^*_t = g, T = t) - E(Y + \delta^*_{Dt}|G^*_t = g, T = t-1)}{E(D|G^*_t = g, T = t) - E(D|G^*_t = g, T = t-1)}, \\
W^*_{CIC}(g,t) =& \frac{E(Y|G^*_t = g, T = t) - E(Q^*_{Dt}(Y)|G^*_t = g, T = t-1)}{E(D|G^*_t = g, T = t) - E(D|G^*_t = g, T = t-1)}.
\end{aligned}$$

When $P(G^*_t = g) = 0$, the three ratios above are not defined. Then, we simply let $W^*_{DID}(g,t) = W^*_{TC}(g,t) = W^*_{CIC}(g,t) = 0$.

Let us then introduce the following weights:

$$\begin{aligned}
w_t &= \frac{DID^*_D(1,t)P(G^*_t = 1, T = t) - DID^*_D(-1,t)P(G^*_t = -1, T = t)}{\sum_{t=1}^{\bar{t}} DID^*_D(1,t)P(G^*_t = 1, T = t) - DID^*_D(-1,t)P(G^*_t = -1, T = t)}, \\
w_{10|t} &= \frac{DID^*_D(1,t)P(G^*_t = 1, T = t)}{DID^*_D(1,t)P(G^*_t = 1, T = t) - DID^*_D(-1,t)P(G^*_t = -1, T = t)},
\end{aligned}$$

where again, we set $DID^*_D(g,t) = 0$ when $P(G^*_t = g) = 0$. The extensions of the Wald-DID, Wald-TC

[5] Using a recategorized treatment $\widetilde{D} = h(D)$ may help alleviating this issue, by weakening the support condition to Supp$(X_{\tilde{d}gt}) =$ Supp(X) for all $\tilde{d} \in$ Supp$(\widetilde{D})$.

and Wald-CIC to multiple groups and periods are defined as

$$W^*_{DID} = \sum_{t\in\mathcal{T}} w_t \left[w_{10|t} W^*_{DID}(1,t) + (1-w_{10|t}) W^*_{DID}(-1,t)\right],$$
$$W^*_{TC} = \sum_{t\in\mathcal{T}} w_t \left[w_{10|t} W^*_{TC}(1,t) + (1-w_{10|t}) W^*_{TC}(-1,t)\right],$$
$$W^*_{CIC} = \sum_{t\in\mathcal{T}} w_t \left[w_{10|t} W^*_{CIC}(1,t) + (1-w_{10|t}) W^*_{CIC}(-1,t)\right].$$

Finally, we consider the following assumption, which replaces Assumption 6.2.

Assumption 6.8. *(Existence of "stable" groups and independence between groups and time)* $\mathcal{T} \neq \emptyset$, *Supp*$(D|G^*_t \neq 0, T = t-1) \subset$ *Supp*$(D|G^*_t = 0, T = t-1)$ *for all* $t \in \mathcal{T}$, *and* $G \perp\!\!\!\perp T$.

Theorem 6.2 below shows that under our previous conditions plus Assumption 6.8, the three estimands point identify Δ. This theorem is proved for the Wald-DID and Wald-TC in [52], and can be proved along the same lines for the Wald-CIC.[6]

Theorem 6.2. *Suppose that Assumptions 6.3 and 6.8 hold.*

1. *If Assumptions 6.4 and 6.5 are satisfied,* $W^*_{DID} = \Delta$.
2. *If Assumption 6.4 ' is satisfied,* $W^*_{TC} = \Delta$.
3. *If Assumptions 6.6 and 6.7 are satisfied,* $W^*_{CIC} = \Delta$.

To estimate W^*_{DID}, W^*_{TC}, and W^*_{CIC}, we suppose that the $(G^*_t)_{t=1\ldots\bar{t}}$ are known. This is the case in applications where the treatment is constant at the group $\times$ period level, as is for instance the case in the example we revisit in Section 6.5. When the $(G^*_t)_{t=1\ldots\bar{t}}$ are unknown, it is also possible to estimate them consistently, without affecting the asymptotic distribution of the estimators of W^*_{DID}, W^*_{TC} and W^*_{CIC}. We refer to Section 2.1 in [54] for details.

Let us focus on the estimator of W^*_{DID}. The estimators of W^*_{TC} and W^*_{CIC} are constructed following exactly the same logic. For any random variable R and any $(g,t) \in \{-1,0,1\} \times \mathcal{T}$, let

$$\widehat{DID}^*_R(g,t) = \frac{1}{n^*_{gt,t}} \sum_{i\in\mathcal{I}^*_{gt,t}} R_i - \frac{1}{n^*_{gt,t-1}} \sum_{i\in\mathcal{I}^*_{gt,t-1}} R_i - \left[\frac{1}{n^*_{0t,t}} \sum_{i\in\mathcal{I}^*_{0t,t}} R_i - \frac{1}{n^*_{0t,t-1}} \sum_{i\in\mathcal{I}^*_{0t,t-1}} R_i\right],$$

where $\mathcal{I}^*_{gt,t'} = \{i : G^*_{ti} = g, T_i = t'\}$ and $n^*_{gt,t'}$ is the size of $\mathcal{I}^*_{gt,t'}$. We let, for $g \in \{-1,0,1\}$, $\widehat{P}(G^*_t = g, T = t) = n^*_{gt,t}/n$. We estimate w_t and $w_{10|t}$ by

$$\widehat{w}_t = \frac{\widehat{DID}^*_D(1,t)\widehat{P}(G^*_t = 1, T = t) - \widehat{DID}^*_D(-1,t)\widehat{P}(G^*_t = -1, T = t)}{\sum_{t=1}^{\bar{t}} \widehat{DID}^*_D(1,t)\widehat{P}(G^*_t = 1, T = t) - \widehat{DID}^*_D(-1,t)\widehat{P}(G^*_t = -1, T = t)},$$
$$\widehat{w}_{10|t} = \frac{\widehat{DID}^*_D(1,t)\widehat{P}(G^*_t = 1, T = t)}{\widehat{DID}^*_D(1,t)\widehat{P}(G^*_t = 1, T = t) - \widehat{DID}^*_D(-1,t)\widehat{P}(G^*_t = -1, T = t)}.$$

We then estimate $W^*_{DID}(g,t)$ by $\widehat{W}^*_{DID}(g,t) = \widehat{DID}^*_Y(g,t)/\widehat{DID}^*_D(g,t)$, and we let

$$\widehat{W}^*_{DID} = \sum_{t\in\mathcal{T}} \widehat{w}_t \left[\widehat{w}_{10|t} \widehat{W}^*_{DID}(1,t) + (1-\widehat{w}_{10|t}) \widehat{W}^*_{DID}(-1,t)\right].$$

[6] [52] obtain the same result on slightly different estimands and without assuming $G \perp\!\!\!\perp T$. Under this additional condition, their estimands are equal to the Wald-DID and Wald-TC considered here. Theorem 6.2 is also similar to Theorem S1 in [54], but they consider slightly different weights and prove the result under stronger conditions.

6.3.3 Other extensions

We now briefly review some other extensions, for which more details can be found in [53] and its supplement.

6.3.3.1 Special cases

When $P(D_{00} = d) = P(D_{01} = d) = 0$ for $d \in \{0, 1\}$, W_{TC} (resp. W_{CIC} and $\tau_{CIC,q}$) is not defined because δ_d (resp. Q_d) is not defined. In such cases, we can simply suppose that $\delta_0 = \delta_1$ (resp. $Q_0 = Q_1$) and modify the estimators accordingly. Then, the Wald-TC becomes equal to the Wald-DID, while the modified CIC estimands identify Δ and τ_q under the same assumptions as above, and if $h_0(h_0^{-1}(y, 1), 0) = h_1(h_1^{-1}(y, 1), 0)$ for every $y \in$ Supp(Y).

6.3.3.2 No "stable" control group

In some applications [see e.g. 68], the treatment rate increases in all groups, thus violating Assumption 6.2. Then, we can still express the Wald-DID as a linear combination of the LATEs of treatment and control group switchers. Specifically, let $S' = \{D(0) \neq D(1), G = 0\}$ be the control group switchers, and $\Delta' = E\left(Y(1) - Y(0)|S', T = 1\right)$ be their local average treatment effect. Under Assumptions 6.1, 6.3, 6.4 and 6.5, we have

$$W_{DID} = \alpha\Delta + (1 - \alpha)\Delta',$$

where $\alpha = (E(D_{11}) - E(D_{10}))/[E(D_{11}) - E(D_{10}) - (E(D_{01}) - E(D_{00}))]$. Hence, the Wald-DID identifies a weighted sum of Δ and Δ'. Note however that if the treatment rate increases in the control group, $E(D_{01}) > E(D_{00})$ and $\alpha > 1$, so Δ' enters with a negative weight. In such a case, we may have $\Delta > 0$ and $\Delta' > 0$ and yet $W_{DID} < 0$. We will only have $W_{DID} = \Delta$ if $\Delta = \Delta'$.

We can also bound Δ under Assumption 6.4 ' if Assumption 6.2 fails. We refer to [53] for such bounds, and to [54] for their corresponding estimators.

6.3.3.3 Non-binary treatment

The Wald-DID, Wald-TC and Wald-CIC still identify a causal parameter if D is not binary but is ordered and takes a finite number of values, as shown in [53]. When the treatment takes a large number of values, its support may differ in the treatment and control groups, and there may be values of D in the treatment group for which δ_d or Q_d are not defined because no unit in the control group has that value of D. This situation includes in particular the special cases discussed above. We can then modify slightly W_{TC} and W_{CIC}. Namely, let us consider a recategorized treatment $\widetilde{D} = h(D)$ grouping together some values of D and let

$$\widetilde{\delta}_{\widetilde{d}} = E[Y_{01}|\widetilde{D} = \widetilde{d}] - E[Y_{00}|\widetilde{D} = \widetilde{d}].$$

We then replace $\delta_{D_{01}}$ by $\widetilde{\delta}_{\widetilde{D}_{01}}$ in the definition of W_{TC}. Then, W_{TC} still identifies Δ provided that $d \mapsto E[Y_{11}(d) - Y_{10}(d)|D(0) = d]$ only depends on $h(d)$. The same applies to W_{CIC}, by using $\widetilde{D}$ instead of D in $Q_d(.)$. Using this recategorized treatment also avoids estimating δ_d and Q_d on a small number of units, thus often lowering the standard errors of the estimators.

Finally, there may also be instances where the treatment has the same support in the treatment and in the control groups, but where bootstrap samples do not satisfy this requirement. For such bootstrap samples, W_{TC} and W_{CIC} cannot be estimated, and the `fuzzydid` command therefore sets them to 10^{15} or -10^{15} with probability $1/2$. To avoid distorting inference, these bootstrap samples are not discarded in

the computation of the percentile-bootstrap confidence intervals, thus enlarging these intervals.[7] This situation is likely to arise when the treatment takes a large number of values. Here again, it may be useful to recategorize the treatment to avoid this issue.

6.4 The `fuzzydid` command

The `fuzzydid` command is compatible with Stata 13.1 and later versions. It uses the `moremata` Stata command to compute estimators with covariates. If this command is not already installed, one must type `ssc install moremata` in Stata's command line.

6.4.1 Syntax

The syntax of `fuzzydid` is as follows:

`fuzzydid Y G T D` [*if*] [*in*] [`, did tc cic lqte newcateg(`*numlist*`) numerator partial nose cluster(`*varname*`) breps(`*#*`) eqtest continuous(`*varlist*`) qualitative(`*varlist*`) modelx(`*reg1 reg2 reg3*`) sieves sieveorder(`*#*`) tagobs`]

6.4.2 Description

`fuzzydid` estimates Δ or τ_q using one or several of the estimators defined in Sections 6.2 and 6.3 above. It also computes their standard errors and confidence intervals.

`Y` is the outcome variable.

`G` is the group variable(s). When the data only bears two groups and two periods, `G` merely corresponds to the variable G defined in Section 6.2, an indicator for units in the treatment group. Outside of this special case, `G` should list the variables G_T^* and G_{T+1}^* defined in Section 6.3.2. We now give an example of a few lines of code that users can follow to create these two variables:

```
sort G T
by G T: egen mean_D = mean(D)
by G: g lag_mean_D = mean_D[_n-1] if G==G[_n-1]&T-1==T[_n-1]
g G_T = sign(mean_D - lag_mean_D)
g G_Tplus1 = G_T[_n+1] if G==G[_n+1]&T+1==T[_n+1]
```

Sometimes, there may not be groups where the treatment is perfectly stable between consecutive periods, thus implying that the Wald-DID, Wald-TC, and Wald-CIC estimators cannot be computed with the `G_T` and `G_Tplus1` variables defined above. Then, the user may replace the 4th line of code above by:

`g G_T = (mean_D - lag_mean_D`$> \varepsilon$`)-(mean_D - lag_mean_D`$< -\varepsilon$`),`

where ε is a positive number small enough to consider that the mean treatment did not really change in groups where it changed by less than ε. See Section 2.1 in [54] for one possible method to choose ε.

`T` is the time period variable, with values in $\{0, ..., \overline{t}\}$.

`D` is the treatment variable. It can be any ordered variable.

[7] They are discarded, on the other hand, in the computation of the bootstrap standard errors.

6.4.3 Options

General options

`did` computes $\widehat{W}_{DID}$ if no covariates are included in the estimation. If some covariates are included, it computes $\widehat{W}^X_{DID,NP}$, $\widehat{W}^X_{DID,OLS}$, or another estimator with covariates depending on the options specified by the user.

`tc` computes $\widehat{W}_{TC}$ if no covariates are included in the estimation. In the special case where D is binary and $P(D_{00} = 0) = P(D_{01} = 0) \in \{0, 1\}$, the command actually computes $\widehat{W}_{DID}$, following the discussion in Section 6.3.3.1. If some covariates are included, it computes $\widehat{W}^X_{TC,NP}$, $\widehat{W}^X_{TC,OLS}$, or another estimator with covariates depending on the options specified by the user.

`cic` computes $\widehat{W}_{CIC}$. In the special case where D is binary and $P(D_{00} = 0) = P(D_{01} = 0) \in \{0, 1\}$, the command actually computes $\widetilde{W}_{CIC}$, following the discussion in Section 6.3.3.1. The `cic` option can only be specified when no covariates are included in the estimation.

`lqte` computes $\widehat{\tau}_q$, for $q \in \{0.05, 0.10, ..., 0.95\}$. This option can only be specified when D, G, and T are binary, and no covariates are included in the estimation. When $P(D_{00} = 0) = P(D_{01} = 0) \in \{0, 1\}$, the command computes $\widetilde{\tau}_{q,CIC}$, following the discussion in Section 6.3.3.1.

At least one of the four options above must be specified. If several of these options are specified, the command computes all the estimators requested by the user.

`newcateg(`*numlist*`)` groups some values of the treatment together when estimating δ_d and Q_d. This option may be useful when the treatment takes a large number of values, as explained in Section 6.3.3.3. The user needs to specify the upper bound of each set of values of the treatment she wants to group. For instance, if D takes the values $\{0, 1, 2, 3, 4.5, 7, 8\}$, and she wants to group together units with $D = \{0, 1, 2\}$, $\{3, 4.5\}$, and $\{7, 8\}$ when estimating δ_d and Q_d, she needs to write `newcateg(2 4.5 8)`.

`numerator` computes only the numerators of the $\widehat{W}_{DID}$, $\widehat{W}_{TC}$ and $\widehat{W}_{CIC}$ estimators. As explained in Section 3.3.3 in [53], this option is useful to conduct placebo tests of the assumptions underlying each estimator.

`partial` computes the bounds of Δ defined in Section 6.3.3.2, $\widehat{\underline{W}}_{TC}$ and $\widehat{\overline{W}}_{TC}$. This option can only be specified when no covariates are included in the estimation.

`nose` computes only the estimators, not their standard errors.

`cluster(`*varname*`)` computes the standard errors of the estimators using a block bootstrap at the *varname* level. Only one clustering variable is allowed.

`breps(#)` specifies the number of bootstrap replications. The default is 50.

`eqtest` performs an equality test between the estimands, when the user specifies at least two of the `did`, `tc`, and `cic` options.

`tagobs` creates a new variable named *tagobs* which identifies the observations used by `fuzzydid`.

Options specific to estimators with covariates

`continuous(`*varlist*`)` specifies the names of all the continuous covariates that need to be included in the estimation.

`qualitative(`*varlist*`)` specifies the names of all the qualitative covariates that need to be included in the estimation. For each variable, indicator variables are created for each value except one, and included as controls in the estimation.

`modelx(`*reg1 reg2 reg3*`)` specifies which parametric method should be used to estimate the conditional expectations in W^X_{DID} or W^X_{TC}. *reg1* specifies which method should be used to estimate $E(Y_{gt}|X)$ and $E(Y_{dgt}|X)$. *reg2* specifies which method should be used to estimate $E(D_{gt}|X)$. When D is not binary,

reg3 specifies which method should be used to estimate $\{P(D_{gt} = d|X)\}_{d\in\{1,...,\overline{d}\}}$. The possible methods are: `ols`, `logit`, and `probit`. For instance, if the user writes `modelx(ols logit logit)`, the command estimates $E(Y_{gt}|X)$ and $E(Y_{dgt}|X)$ by OLS, and $E(D_{gt}|X)$ and $\{P(D_{gt} = d|X)\}_{d\in\{1,...,\overline{d}\}}$ by a logistic regession. The `logit` and `probit` options can only be used with binary variables.

`sieves` indicates that the conditional expectations in W^X_{DID} and W^X_{TC} should be estimated nonparametrically (see Section 6.3.1 above).

When covariates are included in the estimation, and neither `modelx` nor `sieves` is specified, the command estimates by default all conditional expectations by OLS.

`sieveorder(#)` specifies the order of the sieve basis, when the option `sieves` is used. It must be greater than or equal to 2. For a given order L, the number of basis functions is given by $\binom{p_c+L}{L}$ where p_c is the number of continuous covariates. The command does not allow for more than $\min\{4800, n/5\}$ basis functions, where n is the number of observations. If this option is not specified, the choice of the sieve order is done via 5-fold cross-validation with a mean squared error loss function.

6.4.4 Saved results

The `fuzzydid` command saves the following in `e()`:

1. `e(N)`, a scalar containing the number of observations used in the estimation.

2. If the user specifies at least one of the `did`, `tc`, and `cic` options, `fuzzydid` saves `e(b_LATE)`, a $k \times 1$ matrix, where k is equal to the number of options specified. The lines of the matrix correspond to each of the requested estimators. If `nose` is not specified, `fuzzydid` also saves `e(se_LATE)` and `e(ci_LATE)`, a $k \times 1$ and a $k \times 2$ matrix respectively. The lines of `e(se_LATE)` correspond to the bootstrap standard error associated to each of the requested estimators. The columns of `e(ci_LATE)` respectively store the lower and upper bounds of the 95% confidence interval computed by percentile bootstrap for each requested estimator.

3. If the user specifies the `eqtest` option together with at least two of the `did`, `tc`, and `cic` options, `fuzzydid` saves three matrices `e(b_LATE_eqtest)`, `e(se_LATE_eqtest)` and `e(ci_LATE_eqtest)`. The first two matrices have dimension $\binom{k}{2} \times 1$ while the third has dimension $\binom{k}{2} \times 2$, where k is equal to the number of the `did`, `tc`, and `cic` options specified. The matrices `e(b_LATE_eqtest)` and `e(se_LATE_eqtest)` store respectively the value of the difference between each pair of estimators, and the associated bootstrap standard error. The columns of `e(ci_LATE_eqtest)` respectively store the lower and upper bounds of the 95% confidence interval computed by percentile bootstrap associated to each difference.

4. If the user specifies the `lqte` option, the command saves `e(b_LQTE)`, a 19×1 matrix. The lines of the matrix store the value of $\widehat{\tau}_q$ for $q \in \{0.05, 0.10, ..., 0.95\}$. If `nose` is not specified, `fuzzydid` also saves `e(se_LQTE)` and `e(ci_LQTE)`, a 19×1 and a 19×2 matrix respectively. The lines of `e(se_LQTE)` correspond to the bootstrap standard error associated to $\widehat{\tau}_q$ for $q \in \{0.05, 0.10, ..., 0.95\}$. The columns of `e(ci_LQTE)` respectively store the lower and upper bounds of the 95% confidence interval computed by percentile bootstrap for each of the 19 LQTE estimators.

6.5 Example

To illustrate the use of `fuzzydid`, we use the same dataset as [76] to study the effect of newspapers on electoral participation.

`turnout_dailies_1868-1928.dta` is a county-level data set. It contains two variables of interest, `pres_turnout` and `numdailies`, that respectively represent the turnout (Y) and the number of newspapers available (D) in each US county and at each presidential election from 1868 and 1928. First, we load the dataset and present summary statistics:

```
. sum pres_turnout numdailies
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
pres_turnout |     16,872      .65014    .2210102   .0017981      2.518
  numdailies |     16,872    1.463134    2.210448          0         45
```

The average turnout in the 1868 to 1928 presidential elections across counties is 65.01%. The number of newspapers ranges from 0 to 45, and is on average equal to 1.46.

Second, we use `fuzzydid` to compute $\widehat{W}^*_{DID}$, $\widehat{W}^*_{TC}$, and $\widehat{W}^*_{CIC}$ using the first two time periods in the data set, the 1868 and 1872 elections. We then define the `G1872` variable, which is equal to 1 (resp. 0) in counties whose number of newspapers increased (resp. remained stable) between the 1868 and 1872 elections. For now, counties where that number decreased are excluded from the analysis. `numdailies` takes many values, so there are values taken by counties with `G1872`=1 that are not taken by any county with `G1872`=0. Therefore, we use `newcateg` to recategorize `numdailies` into four categories: 0, 1, 2, and 3 or more newspapers.[8] Finally, we cluster the bootstrap at the county level, to allow for county-level correlation over time.

```
. gen G1872=(fd_numdailies>0) if (year==1872)&fd_numdailies!=.&fd_numdailies>=0&
> sample==1
. sort cnty90 year
. replace G1872=G1872[_n+1] if cnty90==cnty90[_n+1]&year==1868
. fuzzydid pres_turnout G1872 year numdailies, did tc cic newcateg(0 1 2 45) bre
> ps(200) cluster(cnty90)
Estimator(s) of the local average treatment effect with bootstrapped standard
errors.  Cluster variable: cnty90.  Number of observations:  1424 .

             |      LATE    Std_Err          t    p_value   lower_ic   upper_ic
-------------+------------------------------------------------------------------
       W_DID |   .0047699   .0160903   .2964428    .766892  -.0230387   .0377381
        W_TC |   .0266618   .0164816   1.617671   .1057335  -.0021458   .0586236
       W_CIC |   .0133223   .0132744   1.003613   .3155653  -.0116416   .0348834
```

The columns of the output table respectively show the value of each estimator, its bootstrap standard error, its t-statistic, its p-value, and the lower and upper bounds of its 95% confidence interval. All point estimates are positive, but none are statistically significant, presumably because this restricted sample with two time periods is too small. In this simple example with two periods and no controls, the computation of the estimators and of 200 bootstrap replications only takes about 3 seconds on a Dell Optiplex 9020 with an Intel Core i7-4790 CPU 3.60 GHz processor and 16GB of RAM, using Stata-MP with four cores.

Third, we compute estimators of the LQTEs, using again the 1868 and 1872 elections. We use a binary treatment variable `numdailies_bin` (0 newspaper, 1 or more), because LQTEs can only be estimated with a binary treatment.

```
. fuzzydid pres_turnout G1872 year numdailies_bin, lqte breps(200) cluster(cnty9
> 0)
```

[8] Only 17.8% of observations have 3 or more newspapers. Results do not change much if instead we recategorize `numdailies` into five categories: 0, 1, 2, 3, and 4 or more newspapers.

```
Estimators of local quantile treatment effects with bootstrapped standard
errors.  Cluster variable: cnty90.  Number of observations:  1424 .
```

	LQTE	Std_Err	t	p_value	lower_ic	upper_ic
q_20	.005	.063113	.0792229	.9368553	-.0825	.1655
q_40	-.052	.0493409	-1.053894	.2919316	-.1244999	.0675
q_60	.011	.0482445	.2280046	.8196427	-.0995	.08
q_80	.02	.0355669	.5623207	.5738975	-.087	.077

To preserve space, we only report $\widehat{\tau}_{0.2}$, $\widehat{\tau}_{0.4}$, $\widehat{\tau}_{0.6}$, and $\widehat{\tau}_{0.8}$, but the command computes $\widehat{\tau}_q$ for $q \in \{0.05, 0.10, ..., 0.95\}$. $\widehat{\tau}_{0.4}$ is negative while the other estimates are positive, thus suggesting that `numdailies_bin` may have heterogeneous effects along the distribution of the outcome. However, none of the point estimates are statistically significant.

Fourth, we compute $\widehat{W}^*_{DID}$, $\widehat{W}^*_{TC}$, and $\widehat{W}^*_{CIC}$ on the full sample. On that purpose, we define the `G_T` and `G_Tplus1` variables described in Section 6.4.2. `G_T` is equal to 1 (resp. 0, -1) for county $c \times$ election-year t observations such that the number of newspapers increased (resp. remained stable, decreased) between election-years $t-1$ and t in that county. `G_Tplus1` is the lead of `G_T`. We add the `eqtest` option, to test whether the estimators are significantly different.

```
. sort cnty90 year
. by cnty90 year: egen mean_D = mean(numdailies)
. by cnty90: g lag_mean_D = mean_D[_n-1] if cnty90==cnty90[_n-1]&year-4==year[_n
> -1]
. g G_T = sign(mean_D - lag_mean_D) if sample==1
. g G_Tplus1 = G_T[_n+1] if cnty90==cnty90[_n+1]&year+4==year[_n+1]
. fuzzydid pres_turnout G_T G_Tplus1 year numdailies, did tc cic newcateg(0 1 2
> 45) breps(200) cluster(cnty90) eqtest
Estimator(s) of the local average treatment effect with bootstrapped standard
errors.  Cluster variable: cnty90.  Number of observations:  16872 .
```

	LATE	Std_Err	t	p_value	lower_ic	upper_ic
W_DID	.0037507	.0012813	2.927357	.0034186	.0009971	.0057828
W_TC	.0053305	.0013276	4.015155	.0000594	.0023461	.0075914
W_CIC	.004215	.001477	2.853841	.0043194	.0009549	.0067769

```
Estimators equality test
```

	Delta	Std_Err	t	p_value	lower_ic	upper_ic
DID_TC	-.0015798	.0003504	-4.507975	6.54e-06	-.0023752	-.0009441
DID_CIC	-.0004643	.0007151	-.6492892	.5161515	-.0018629	.0008515
TC_CIC	.0011155	.0006505	1.71487	.086369	-.0002291	.0023088

The Wald-DID is equal to 0.0038. According to that estimator, increasing the number of newspapers available in a county by one increases voters' turnout in presidential elections by 0.38 percentage points. This estimator is significantly different from 0 at the 5% level. The Wald-TC is larger (0.0053), and significantly different from the Wald-DID (t-stat=-4.51). The Wald-CIC lies in between (0.0042), and this estimator is not significantly different from the other two. In this more complicated example with 16 periods and almost 17 000 observations, the computation of the estimators and of 200 bootstrap replications still only takes around two minutes.

[76] allow for state-specific trends in their specification, so we compute $\widehat{W}^*_{DID}$ and $\widehat{W}^*_{TC}$ with state indicators as controls, which is equivalent to allowing for state-specific trends.[9]

```
. fuzzydid pres_turnout G_T G_Tplus1 year numdailies, did tc newcateg(0 1 2 45)
> qualitative(st1-st48) breps(200) cluster(cnty90) eqtest
```

[9] On the other hand, `fuzzydid` does not compute $\widehat{W}^*_{CIC}$ with controls.

```
Estimator(s) of the local average treatment effect with bootstrapped standard
errors.  Cluster variable: cnty90.  Number of observations:  16872 .  Controls
included in the estimation:  st1 st2 st3 st4 st5 st6 st7 st8 st9 st10 st11
st12 st13 st14 st15 st16 st17 st18 st19 st20 st21 st22 st23 st24 st25 st26
st27 st28 st29 st30 st31 st32 st33 st34 st35 st36 st37 st38 st39 st40 st41
st42 st43 st44 st45 st46 st47 st48 .
```

	LATE	Std_Err	t	p_value	lower_ic	upper_ic
W_DID	.0026383	.0012213	2.160195	.0307575	.0002316	.0048236
W_TC	.0043428	.0014116	3.076507	.0020944	.0015519	.0066773

```
Estimators equality test
```

	Delta	Std_Err	t	p_value	lower_ic	upper_ic
DID_TC	-.0017046	.0009193	-1.85417	.0637148	-.0034308	.0000123

With those controls, $\widehat{W}^*_{DID} = 0.0026$ and $\widehat{W}^*_{TC} = 0.0043$, and the two estimators are significantly different at the 10% level (t-stat=-1.85). Adding the control variables substantially increases the computation time, to 79 minutes.

Finally, we compute a placebo Wald-DID (resp. Wald-TC) estimator, to assess if Assumptions 6.4 and 6.5 (resp. Assumption 6.4 ') are plausible in this application. Instead of using the turnout in county g and election-year t as the outcome variable, our placebo estimators use the turnout in the same county in the previous election. Moreover, only counties where the number of newspapers did not change between $t-2$ and $t-1$ are included in the estimation. Therefore, our placebo estimators compare the evolution of turnout from $t-2$ to $t-1$, between counties where the number of newspapers increased or decreased between $t-1$ and t and counties where that number remained stable, restricting the sample to counties where the number of newspapers remained stable from $t-2$ to $t-1$.

```
. xtset cnty90 year
. gen fd_numdailies_l1=l4.fd_numdailies
. gen pres_turnout_l1=l4.pres_turnout
. sort cnty90 year
. g G_T_placebo = sign(mean_D - lag_mean_D) if sample==1&fd_numdailies_l1==0
. g G_Tplus1_placebo = G_T_placebo[_n+1] if cnty90==cnty90[_n+1]&year+4==year[_n
> +1]
. fuzzydid pres_turnout_l1 G_T_placebo G_Tplus1_placebo year numdailies, did tc
> newcateg(0 1 2 45) qualitative(st1-st48) breps(200) cluster(cnty90)
Estimator(s) of the local average treatment effect with bootstrapped standard
errors.  Cluster variable: cnty90.  Number of observations:  13221 .  Controls
included in the estimation:  st1 st2 st3 st4 st5 st6 st7 st8 st9 st10 st11
st12 st13 st14 st15 st16 st17 st18 st19 st20 st21 st22 st23 st24 st25 st26
st27 st28 st29 st30 st31 st32 st33 st34 st35 st36 st37 st38 st39 st40 st41
st42 st43 st44 st45 st46 st47 st48 .
```

	LATE	Std_Err	t	p_value	lower_ic	upper_ic
W_DID	-.00183	.0016594	-1.102842	.2700959	-.0051247	.0013008
W_TC	-.0008691	.0018412	-.4720226	.6369107	-.0041261	.0025142

The placebo Wald-DID is negative, indicating that the actual Wald-DID may be downward biased due to a violation of Assumptions 6.4 and 6.5. However, this placebo estimator is not statistically significant. The placebo Wald-TC is also negative and not statistically significant. It is twice smaller than the placebo Wald-DID, thus indicating that Assumption 6.4 ' may be more plausible than Assumptions 6.4 and 6.5 in this application.

6.6 Monte Carlo Simulations

This section exhibits the finite sample performance of the estimators of W_{DID}, W_{TC}, W_{CIC} and $\tau_{CIC,q}$. We consider for that purpose the folllowing DGP. Let (G,T) be uniform on $\{0,1\}^2$. Let $(U(0),U(1),V) \sim \mathcal{N}(0,\Sigma)$, with $\Sigma_{ii}=1$ for $i \in \{1,3\}$, $\Sigma_{22}=1.2$, $\Sigma_{12}=0$, $\Sigma_{13}=.5$ and $\Sigma_{23}=-.5$, and with $(U(0),U(1),V) \perp\!\!\!\perp (G,T)$. Then we let

$$Y(d) = d + G + T + U(d),$$
$$D(t) = \mathbb{1}\{V \geq 1 - G \times t\}.$$

In this DGP, all the assumptions in Section 6.2 hold. Therefore, W_{DID}, W_{TC}, and W_{CIC} all identify Δ, while $\tau_{CIC,q}$ identifies τ_q. We focus on the bias, mean square error, and coverage rate of estimators of Δ and τ_q for $q \in \{.25,.5,.75\}$, and for sample sizes equal to 400, 800, and 1,600. In this DGP, $\Delta \simeq .540$, $\tau_{.25} \simeq .481$, $\tau_{.5} \simeq .536$ and $\tau_{.75} \simeq .595$.

The results are displayed in Table 6.1. Even with small samples, the Wald-DID and Wald-TC estimators do not exhibit any systematic bias. Their RMSE are also very similar. The Wald-CIC, on the other hand, is more biased and has a RMSE which is 5 to 15% larger. This is probably due to the estimator of the nonlinear transform Q_d. This estimator is likely biased and imprecise in the tails, which may also explain the bias and high RMSE of $\widehat{\tau}_q$ for $n = 400$. Note however that the bias of $\widehat{W}_{CIC}$, $\widehat{\tau}_{.25}$, $\widehat{\tau}_{.5}$, and $\widehat{\tau}_{.75}$ decreases quickly with the sample size. For $n = 1,600$, the bias of these estimators is already negligible compared to their RMSE. Finally, the percentile bootstrap confidence intervals of all estimators are quite accurate, with all coverage rates lying between .92 and .97 when the nominal level is .95. The levels are slightly more distorted for the Wald-CIC and the $\widehat{\tau}_q$ but again, they become closer to 95% as the sample size increases.

Table 6.1 – Results of the Monte Carlo simulations

		Estimators of Δ			Estimators of τ_q		
n	Statistic	$\widehat{W}_{DID}$	$\widehat{W}_{TC}$	$\widehat{W}_{CIC}$	$\widehat{\tau}_{.25}$	$\widehat{\tau}_{.5}$	$\widehat{\tau}_{.75}$
400	Bias	0,005	-0,002	0,174	0,002	-0,154	-0,497
	RMSE	0,651	0,613	0,682	0,712	0,867	1,223
	Cov. rate	0,948	0,948	0,921	0,971	0,967	0,917
800	Bias	0,015	0,01	0,088	-0,056	-0,029	-0,235
	RMSE	0,422	0,414	0,472	0,539	0,555	0,922
	Cov. rate	0,953	0,951	0,929	0,964	0,961	0,934
1600	Bias	-0,005	-0,005	0,034	-0,054	-0,013	-0,077
	RMSE	0,286	0,284	0,329	0,394	0,382	0,58
	Cov. rate	0,948	0,946	0,943	0,964	0,966	0,955

Notes: "Cov. rate" stands for coverage rates of (percentile bootstrap) confidence intervals, with a nominal level of 95%. The results are based on 1,000 samples and for each, 500 bootstrap samples are drawn to construct the confidence intervals. With our DGP, $\Delta \simeq .540$, $\tau_{.25} \simeq .481$, $\tau_{.5} \simeq .536$ and $\tau_{.75} \simeq .595$.

6.7 Conclusion

We have discussed how to use `fuzzydid` to estimate local average and quantile treatment effects in fuzzy differences-in-differences designs, following de Chaisemartin and D'Haultfœuille [53]. In such designs, the popular Wald-DID estimand relies on a stable treatment effect assumption, which may not be plausible. Then, the Wald-TC and Wald-CIC estimands may be valuable alternatives, as they do not hinge upon this assumption. Similarly, when the data bears multiple groups and periods, the Wald-TC and Wald-CIC estimands may be valuable alternatives to commonly used two-way linear regressions. The `fuzzydid` command makes it easy to estimate those estimands.

Bibliography

[1] Daron Acemoglu, David Autor, David Dorn, Gordon H. Hanson, and Brendan Price. Import competition and the great us employment sag of the 2000s. *Journal of Labor Economics*, 34(S1): S141–S198, 2016.

[2] Daniel Ackerberg, Xiaohong Chen, Jinyong Hahn, and Zhipeng Liao. Asymptotic Efficiency of Semiparametric Two-step GMM. *The Review of Economic Studies*, 81(3):919–943, 2014.

[3] Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

[4] David J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):pp. 581–598, 1981.

[5] Pierre Alquier, Vincent Cottet, and Guillaume Lecué. Estimation bounds and sharp oracle inequalities of regularized procedures with lipschitz loss functions. *Ann. Statist.*, 47(4):2117–2144, 08 2019.

[6] Theodore W Anderson and Herman Rubin. Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20(1):46–63, 1949.

[7] Donald W.K. Andrews. Examples of l2-complete and boundedly-complete distributions. *Journal of Econometrics*, 199(2):213 – 220, 2017.

[8] Isaiah Andrews, James Stock, and Liyang Sun. Weak instruments in IV regression: Theory and practice. *To appear in Annual Review of Economics*, 2019.

[9] Miguel Arcones and Evarist Giné. Limit theorems for U-processes. *The Annals of Probability*, 21 (3):pp. 1494–1542, 1993.

[10] Timothy Armstrong and Michal Kolesár. Simple and honest confidence intervals in nonparametric regression. *arXiv preprint arXiv:1606.01200*, 2019.

[11] Timothy Armstrong and Michal Kolesár. Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683, 2018.

[12] Susan Athey and Guido Imbens. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497, 2006.

[13] Andrii Babii and Jean-Pierre Florens. Distribution of residuals in the nonparametric iv model with application to separability testing. Technical report, Toulouse School of Economics, 2017.

[14] Raghu R Bahadur and Leonard J Savage. The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics*, 27(4):1115–1122, 1956.

[15] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. *Inference for High-Dimensional Sparse Econometric Models*, volume 3 of *Econometric Society Monographs*, page 245–295. Cambridge University Press, 2013.

[16] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50, May 2014.

[17] Alexandre Belloni, Victor Chernozhukov, Iv'an Fernández-Val, and Christian Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

[18] Patrice Bertail, Emilie Chautru, and Stephan Clémencon. Empirical processes in survey sampling with (conditional) poisson designs. *Scandinavian Journal of Statistics*, 44(1):97–111, 2017.

[19] Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1):249–275, 2004.

[20] Nicolai Bissantz, Thorsten Hohage, and Axel Munk. Consistency and rates of convergence of nonlinear tikhonov regularization with random noise. *Inverse Problems*, 20(6):1773–1789, 2004.

[21] Richard Blundell, Martin Browning, and Ian Crawford. Nonparametric engel curves and revealed preference. *Econometrica*, 71(1):205–240, 2003.

[22] Richard Blundell, Xiaohong Chen, and Dennis Kristensen. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669, 2007.

[23] Richard Blundell, Martin Browning, and Ian Crawford. Best nonparametric bounds on demand responses. *Econometrica*, 76(6):1227–1262, 2008.

[24] Jonathan Borwein and Adrian Lewis. Duality relationships for entropy-like minimization problems. *SIAM Journal on Control and Optimization*, 29(2):325–338, 1991.

[25] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[26] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[27] Anthony Briant, Miren Lafourcade, and Benoît Schmutz. Can tax breaks beat geography? lessons from the french enterprise zone experience. *American Economic Journal: Economic Policy*, 7(2): 88–124, May 2015.

[28] Axel Bücher and Ivan Kojadinovic. A note on conditional versus joint unconditional weak convergence in bootstrap consistency results. *To appear in Journal of Theoretical Probability*, 2019.

[29] T. Tony Cai and Mark Low. Adaptive confidence balls. *Ann. Statist.*, 34(1):202–228, 02 2006.

[30] Colin Cameron, Jonah Gelbach, and Douglas Miller. Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2):238–249, 2011.

[31] Marine Carrasco and Rachidi Kotchoni. Regularized generalized empirical likelihood estimators. Technical report, 2017.

[32] Marine Carrasco, Jean-Pierre Florens, and Eric Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. volume 6, Part B of *Handbook of Econometrics*, pages 5633 – 5751. Elsevier, 2007.

[33] Marine Carrasco, Jean-Pierre Florens, and Eric Renault. Asymptotic normal inference in linear inverse problem. volume 73 of *Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*. Oxford University Press, 2013.

[34] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012.

[35] Laurent Cavalier. Inverse problems in statistics. In *Inverse problems and high-dimensional estimation: stats in the Château Summer School, August 31-September 4, 2009*, volume 203. Springer Science & Business Media, 2011.

[36] Xiaohong Chen and Timothy Christensen. Optimal sup-norm rates, adaptivity and inference in nonparametric instrumental variables estimation. Technical report, Cowles Foundation, 2015.

[37] Xiaohong Chen and Demian Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321, 2012.

[38] Xiaohong Chen and Demian Pouzo. Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica*, 83(3):1013–1079, 2015.

[39] Xiaohong Chen and Markus Reiss. On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory*, 27(3):497–521, 2011.

[40] Xiaohong Chen, Demian Pouzo, and James Powell. Penalized sieve gel for weighted average derivatives of nonparametric quantile iv regressions. Technical report, 2019.

[41] Victor Chernozhukov, Guido Imbens, and Whitney Newey. Instrumental variable estimation of nonseparable models. *Journal of Econometrics*, 139(1):4 – 14, 2007.

[42] Victor Chernozhukov, Iván Fernández-Val, and Alfred Galichon. Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125, 2010.

[43] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximation of suprema of empirical processes. *Annals of Statistics*, 42(4):1564–1597, 2014.

[44] Denis Chetverikov and Daniel Wilhelm. Nonparametric instrumental variable estimation under monotonicity. *Econometrica*, 85(4):1303–1320, 2017.

[45] Geoffrey Chinot, Guillaume Lecu'e, and Matthieu Lerasle. Statistical learning with lipschitz and convex loss functions. *ArXiv preprint, arXiv:1810.01090*, 2019.

[46] Eve Colson-Sihra and Clement Bellet. The conspicuous consumption of the poor: Forgoing calories for aspirational goods. *SSRN Electronic Journal*, 01 2018.

[47] Bruno Crépon, Marc Ferracci, and Denis Fougère. Training the unemployed in france: How does it affect unemployment duration and recurrence? *Annals of Economics and Statistics*, (107/108): 175–199, 2012.

[48] Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

[49] Laurent Davezies. On the existence of $\sqrt{N}$-consistent estimators of linear functionals in non-parametric iv models. Technical report, CREST, 2016.

[50] Laurent Davezies, Xavier D'Haultfœuille, and Yannick Guyonvarch. Asymptotic results under multiway clustering. ArXiv e-prints, eprint 1807.07925, 2018.

[51] Laurent Davezies, Xavier D'Haultfœuille, and Yannick Guyonvarch. Empirical process results for exchangeable arrays. *ArXiv preprint, arXiv:1906.11293*, 2019.

[52] Clément de Chaisemartin and Xavier D'Haultfœuille. Two-way fixed effects estimators with heterogeneous treatment effects. Working paper, 2018.

[53] Clément de Chaisemartin and Xavier D'Haultfœuille. Fuzzy differences-in-differences. *Review of Economic Studies*, 85(2):999–1028, 2018.

[54] Clément de Chaisemartin and Xavier D'Haultfœuille. Supplement to "fuzzy differences-in-differences". *Review of Economic Studies Supplementary Material*, 85(2), 2018.

[55] Victor de la Peña. Decoupling and khintchine's inequalities for u-statistics. *The Annals of Probability*, pages 1877–1892, 1992.

[56] Victor de la Peña and Evarist Giné. *Decoupling. Probability and its Applications*. Springer-Verlag, New York, 1999.

[57] Herold Dehling and Walter Philipp. *Empirical process techniques for dependent data*. Springer, 2002.

[58] Alexis Derumigny, Lucas Girard, and Yannick Guyonvarch. On the construction of confidence intervals for ratios of expectations. *ArXiv preprint, arXiv:1904.07111*, 2019.

[59] Luc Devroye. The uniform convergence of the nadaraya-watson regression function estimate. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 6(2):179–191, 1978.

[60] Xavier D'Haultfoeuille. *Essays on some Identification Issues in Economics*. PhD thesis, Université Paris 1 Panthéon-Sorbonne, 2009.

[61] Xavier D'Haultfoeuille. On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27(3):460–471, 2011.

[62] Xavier D'Haultfoeuille. Lecture notes on ivs in nonparametric/nonlinear models, 2019.

[63] Jean-Marie Dufour. Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica*, 65(6):1365–1387, 1997.

[64] Fabian Dunker, Jean-Pierre Florens, Thorsten Hohage, Jan Johannes, and Enno Mammen. Iterative estimation of solutions to noisy nonlinear operator equations in nonparametric instrumental regression. *Journal of Econometrics*, 178:444 – 455, 2014.

[65] Xavier D'Haultfœuille and Pauline Givord. La régression quantile en pratique. *Economie et Statistique*, 471(1):85–111, 2014.

[66] G. K. Eagleson and Neville Weber. Limit theorems for weakly exchangeable arrays. *Mathematical Proceedings of the Cambridge Philosophical Society*, 84(1):123–130, 1978.

[67] Bradley Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 01 1979.

[68] Ruben Enikolopov, Maria Petrova, and Ekaterina Zhuravskaya. Media and political persuasion: Evidence from russia. *The American Economic Review*, 101:3253, 2011.

[69] Marcel Fafchamps and Flore Gubert. The formation of risk sharing networks. *Journal of development Economics*, 83(2):326–350, 2007.

[70] Jianqing Fan. Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.*, 21 (1):196–216, 03 1993.

[71] Max Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. *arXiv preprint arXiv:1809.09953*, 2018.

[72] Joachim Freyberger and Matthew Masten. A practical guide to compact infinite dimensional parameter spaces. *Accepted at Econometric Reviews*, 2018.

[73] Bert Fristedt and Lawrence Gray. *A Modern Approach to Probability Theory*. Probability and Its Applications. Birkhäuser Boston, 2013.

[74] Markus Frölich. Nonparametric iv estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139:35–75, 2007.

[75] Theo Gasser and Hans-Georg Müller. Kernel estimation of regression functions. In Th. Gasser and M. Rosenblatt, editors, *Smoothing Techniques for Curve Estimation*, pages 23–68. Springer Berlin Heidelberg, 1979.

[76] Matthew Gentzkow, Jesse M Shapiro, and Michael Sinkinson. The effect of newspaper entry and exit on electoral politics. *The American Economic Review*, 101:2980, 2011.

[77] Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(6):907 – 921, 2002.

[78] Evarist Giné and David Mason. On local u -statistic processes and the estimation of densities of functions of several sample variables. *The Annals of Statistics*, 35(3):1105–1145, 07 2007.

[79] Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015.

[80] Evarist Giné, Vladimir Koltchinskii, and Jon Wellner. Ratio limit theorems for empirical processes. In Evariste Giné, Christian Houdré, and David Nualart, editors, *Stochastic Inequalities and Applications*, pages 249–278. Birkhäuser Basel, 2003.

[81] Pauline Givord. Méthodes économétriques pour l'évaluation de politiques publiques. *Economie et Prévision*, 204-205(1):1–28, 2014.

[82] Christian Gourieroux, Alain Monfort, and Alain Trognon. Pseudo maximum likelihood methods: applications to poisson models. *Econometrica*, 52:701–720, 1984.

[83] Bryan Graham. Lecture notes on dyadic regression, October 2018.

[84] Allan Gut. Complete convergence for arrays. *Periodica Mathematica Hungarica*, 25(1):51–75, 1992.

[85] Yannick Guyonvarch. Nonparametric estimation in conditional moment restricted models via generalized empirical likelihood. Technical report, CREST, 2019.

[86] Peter Hall and Joel Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33, 05 2005.

[87] Peter Hall and Bing-Yi Jing. Uniform coverage bounds for confidence intervals and berry-esseen theorems for edgeworth expansion. *Ann. Statist.*, 23(2):363–375, 04 1995.

[88] Lars Peter Hansen, John Heaton, and Amir Yaron. Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3):262–280, 1996.

[89] Douglas Hoover. Relations on probability spaces and arrays of random variables. Working paper, 1979.

[90] Joel Horowitz and Sokbae Lee. Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica*, 75(4):1191–1208, 2007.

[91] Joel Horowitz and Sokbae Lee. Nonparametric estimation and inference under shape restrictions. *Journal of Econometrics*, 201(1):108 – 126, 2017.

[92] Guido Imbens and Joshua Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

[93] Guido Imbens, Stephen Donald, and Whitney Newey. Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*, 117(1):55–93, 2003.

[94] Koen Jochmans. Two-way models for gravity. *Review of Economics and Statistics*, 99(3):478–485, 2017.

[95] Olav Kallenberg. On the representation theorem for exchangeable arrays. *Journal of Multivariate Analysis*, 30(1):137–154, 1989.

[96] Olav Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Springer, 2005.

[97] Maximilian Kasy. Uniformity and the delta method. *Journal of Econometric Methods*, 8(1), 2019.

[98] Kengo Kato. Lecture notes on empirical process theory. Technical report, Cornell University, 2019.

[99] Yuichi Kitamura. Empirical Likelihood methods in econometrics: Theory and practice. In Richard Blundell, Whitney Newey, and Torsten Persson, editors, *Advances in Economics and Econometrics*, volume 3, pages 174–237. Cambridge University Press, 2007. Cambridge Books Online.

[100] Yuichi Kitamura and Michael Stutzer. An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65(4):pp. 861–874, 1997.

[101] Yuichi Kitamura, Gautam Tripathi, and Hyungtaik Ahn. Empirical Likelihood-based inference in conditional moment restriction models. *Econometrica*, 72(6):1667–1714, 2004.

[102] Michael Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer Verlag New York, 2006.

[103] Pascal Lavergne and Valentin Patilea. Smooth minimum distance estimation and testing with conditional estimating equations: Uniform in bandwidth theory. *Journal of Econometrics*, 177(1):47 – 59, 2013.

[104] Ker-Chau Li. Honest confidence regions for nonparametric regression. *Ann. Statist.*, 17(3): 1001–1008, 09 1989.

[105] James MacKinnon, Morten Nielsen, and Matthew Webb. Wild bootstrap and asymptotic inference with multiway clustering. Working paper, 2019.

[106] Enno Mammen and Alexandre Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6): 1808–1829, 12 1999.

[107] Elias Masry. Multivariate local polynomial regression for time series:uniform strong consistency and rates. *Journal of Time Series Analysis*, 17(6):571–599, 1996.

[108] Peter McCullagh et al. Resampling and exchangeable arrays. *Bernoulli*, 6(2):285–301, 2000.

[109] Konrad Menzel. Bootstrap with clustering in two or more dimensions. Working paper, 2018.

[110] Enrico Moretti. Real wage inequality. *American Economic Journal: Applied Economics*, 5(1): 65–103, January 2013.

[111] Whitney Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

[112] Whitney Newey and James Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

[113] Whitney Newey and Richard Smith. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

[114] Richard Nickl and Benedikt Pötscher. Bracketing metric entropy rates and empirical central limit theorems for function classes of besov-and sobolev-type. *Journal of Theoretical Probability*, 20(2): 177–199, 2007.

[115] Deborah Nolan and David Pollard. u-processes: Rates of convergence. *The Annals of Statistics*, 15(2):780–799, 06 1987.

[116] Taisuke Otsu. Empirical likelihood estimation of conditional moment restriction models with unknown functions. *Econometric Theory*, 27(1):8–46, 2011.

[117] Art Owen. Empirical Likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2): 237–249, 1988.

[118] Art Owen. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411, 2007.

[119] Paulo Parente and Richard Smith. Gel methods for nonsmooth moment indicators. *Econometric Theory*, 27(1):74–113, 2011.

[120] Iosif Pinelis and Raymond Molzon. Optimal-order bounds on the rate of convergence to normality in the multivariate delta method. *Electronic Journal of Statistics*, 10(1):1001–1063, 2016.

[121] James Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89 (427):846–866, 1994.

[122] Joseph Romano. On non-parametric testing, the uniform behaviour of the t-test, and related problems. *Scandinavian Journal of Statistics*, 31(4):567–584, 2004.

[123] Joseph Romano and Michael Wolf. Finite sample nonparametric inference and large sample efficiency. *The Annals of Statistics*, 28(3):756–778, 2000.

[124] Donald Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[125] Andres Santos. Instrumental variable methods for recovering continuous linear functionals. *Journal of Econometrics*, 161(2):129 – 146, 2011.

[126] João Santos Silva and Silvana Tenreyro. The log of gravity. *The Review of Economics and statistics*, 88(4):641–658, 2006.

[127] Xiaotong Shen and Wing Hung Wong. Convergence rate of sieve estimates. *The Annals of Statistics*, 22(2):580–615, 06 1994.

[128] Bernard Silverman. Limit theorems for dissociated random variables. *Advances in Applied Probability*, 8(4):806–819, 1976.

[129] Douglas Staiger and James Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586, 1997.

[130] James Stock and Motohiro Yogo. *Testing for Weak Instruments in Linear IV Regression*, pages 80–108. Cambridge University Press, New York, 2005.

[131] Max Tabord-Meehan. Inference with dyadic data: Asymptotic behavior of the dyadic-robust t-statistic. *Journal of Business and Economic Statistics*, Forthcoming, 2019.

[132] Terence Tao. *An Introduction to Measure Theory*. Graduate studies in mathematics. American Mathematical Society, 2011.

[133] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[134] Alexander Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1): 135–166, 02 2004.

[135] Alexandre Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2009.

[136] Aad van der Vaart. *Asymptotics Statistics*. Cambridge University Press, 2000.

[137] Aad van der Vaart and Jon Wellner. *Weak Convergence of Empirical Processes: with Applications to Statistics*. Springer-Verlag New York, 1996.

[138] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[139] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences de l'homme et de la société (SHS)

Titre: Contributions à l'estimation et à l'inférence robuste en économétrie semi- et nonparamétrique
Mots clés: variables instrumentales, processus empiriques, échangeabilité

Résumé: Dans le chapitre introductif, nous dressons une étude comparée des approches en économétrie et en apprentissage statistique sur les questions de l'estimation et de l'inférence en statistique.
Dans le deuxième chapitre, nous nous intéressons à une classe générale de modèles de variables instrumentales nonparamétriques. Nous généralisons la procédure d'estimation de [116] en y ajoutant un terme de régularisation. Nous prouvons la convergence de notre estimateur pour la norme L_2 de Lebesgue.
Dans le troisième chapitre, nous montrons que lorsque les données ne sont pas indépendantes et identiquement distribuées (*i.i.d*) mais simplement jointement échangeables, une version modifiée du processus empirique converge faiblement vers un processus gaussien sous les mêmes conditions que dans le cas *i.i.d*. Nous obtenons un résultat similaire pour une version adaptée du processus empirique bootstrap. Nous déduisons de nos résultats la normalité asymptotique de plusieurs estimateurs non-linéaires ainsi que la validité de l'inférence basée sur le bootstrap. Nous revisitons enfin l'article empirique de [126].
Dans le quatrième chapitre, nous abordons la question de l'inférence pour des ratios d'espérances. Nous trouvons que lorsque le dénominateur ne tend pas trop vite vers zéro quand le nombre d'observations n augmente, le bootstrap nonparamétrique est valide pour faire de l'inférence asymptotique. Dans un second temps, nous complétons un résultat d'impossibilité de [63] en montrant que quand n est fini, il est possible de construire des intervalles de confiance qui ne sont pas pathologiques sont certaines conditions sur le dénominateur.
Dans le cinquième chapitre, nous présentons une commande Stata qui implémente les estimateurs proposés par [53] pour mesurer plusieurs types d'effets de traitement très étudiés en pratique.

Title: Essays in robust estimation and inference in semi- and nonparametric econometrics
Keywords: instrumental variables, empirical processes, exchangeability

Abstract: In the introductory chapter, we compare views on estimation and inference in the econometric and statistical learning disciplines.
In the second chapter, our interest lies in a generic class of nonparametric instrumental models. We extend the estimation procedure in [116] by adding a regularisation term to it. We prove the consistency of our estimator under Lebesgue's L_2 norm.
In the third chapter, we show that when observations are jointly exchangeable rather than independent and identically distributed (*i.i.d*), a modified version of the empirical process converges weakly towards a Gaussian process under the same conditions as in the *i.i.d* case. We obtain a similar result for a modified version of the bootstrapped empirical process. We apply our results to get the asymptotic normality of several non-linear estimators and the validity of bootstrap-based inference. Finally, we revisit the empirical work of [126].
In the fourth chapter, we address the issue of conducting inference on ratios of expectations. We find that when the denominator tends to zero slowly enough when the number of observations n increases, bootstrap-based inference is asymptotically valid. Secondly, we complement an impossibility result of [63] by showing that whenever n is finite it is possible to construct confidence intervals which are not pathological under some conditions on the denominator.
In the fifth chapter, we present a Stata command which implements estimators proposed in [53] to measure several types of treatment effects widely studied in practice.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France