Modelling the dependence of order statistics and nonparametric estimation.

Richard Fischer

UNIVERSITÉ PARIS-EST

L'École Doctorale Mathématiques et Sciences et Technologies de l'information et de la Communication (MSTIC)

THÈSE DE DOCTORAT

Discipline : Mathématiques Appliquées

présentée par

Richárd FISCHER

Modélisation de la dépendance pour des statistiques d'ordre et estimation non-paramétrique

Thèse co-dirigée par Cristina BUTUCEA, Jean-François DELMAS et Anne DUTFOY

préparée au CERMICS, Université Paris-Est (ENPC) et au LAMA, Université Paris-Est (UPE-MLV)

Thèse soutenue le 30/09/2016 devant le jury composé de :

Mme Cristina BUTUCEA	Directrice de Thèse	Université Paris-Est Marne-la-Vallée
M. Jean-François DELMAS	Directeur de Thèse	École des Ponts ParisTech
Mme Anne DUTFOY	Directrice de Thèse	EDF R&D
M. Jean-David FERMANIAN	Rapporteur	ENSAE ParisTech
M. Johan SEGERS	Rapporteur	Université Catholique de Louvain
Mme Agnès SULEM	Examinatrice	INRIA
M. Alexandre TSYBAKOV	Examinateur	ENSAE ParisTech

Résumé

Dans cette thèse, on considère la modélisation de la loi jointe des statistiques d'ordre, c.à.d. des vecteurs aléatoires avec des composantes ordonnées presque sûrement. La première partie est dédiée à la modélisation probabiliste des statistiques d'ordre d'entropie maximale à marginales fixées. Les marginales étant fixées, la caractérisation de la loi jointe revient à considérer la copule associée. Dans le Chapitre 2, on présente un résultat auxiliaire sur les copules d'entropie maximale à diagonale fixée. Une condition nécessaire et suffisante est donnée pour l'existence d'une telle copule, ainsi qu'une formule explicite de sa densité et de son entropie. La solution du problème de maximisation d'entropie pour les statistiques d'ordre à marginales fixées est présentée dans le Chapitre 3. On donne des formules explicites pour sa copule et sa densité jointe. On applique le modèle obtenu pour modéliser des paramètres physiques dans le Chapitre 4.

Dans la deuxième partie de la thèse, on étudie le problème d'estimation non-paramétrique des densités d'entropie maximale des statistiques d'ordre en distance de Kullback-Leibler. Le chapitre 5 décrit une méthode d'agrégation pour des densités de probabilité et des densités spectrales, basée sur une combinaison convexe de ses logarithmes, et montre des bornes optimales non-asymptotiques en déviation. Dans le Chapitre 6, on propose une méthode adaptative issue d'un modèle exponentiel log-additif pour estimer les densités considérées, et on démontre qu'elle atteint les vitesses connues minimax. L'application de cette méthode pour estimer des dimensions des défauts est présentée dans le Chapitre 7.

Mots clés

agrégation, copule, densité de probabilité, densité spectrale, divergence de Kullback-Leibler, entropie maximale, estimation adaptative, estimation non-paramétrique, modèle exponentiel log-additif, statistiques d'ordre

Abstract

Modelling the dependence of order statistics and nonparametric estimation

In this thesis we consider the modelling of the joint distribution of order statistics, i.e. random vectors with almost surely ordered components. The first part is dedicated to the probabilistic modelling of order statistics of maximal entropy with marginal constraints. Given the marginal constraints, the characterization of the joint distribution can be given by the associated copula. Chapter 2 presents an auxiliary result giving the maximum entropy copula with a fixed diagonal section. We give a necessary and sufficient condition for its existence, and derive an explicit formula for its density and entropy. Chapter 3 provides the solution for the maximum entropy problem for order statistics with marginal constraints by identifying the copula of the maximum entropy distribution. We give explicit formulas for the copula and the joint density. An application for modelling physical parameters is given in Chapter 4.

In the second part of the thesis, we consider the problem of nonparametric estimation of maximum entropy densities of order statistics in Kullback-Leibler distance. Chapter 5 presents an aggregation method for probability density and spectral density estimation, based on the convex combination of the logarithms of these functions, and gives non-asymptotic bounds on the aggregation rate. In Chapter 6, we propose an adaptive estimation method based on a log-additive exponential model to estimate maximum entropy densities of order statistics which achieves the known minimax convergence rates. The method is applied to estimating flaw dimensions in Chapter 7.

Keywords

adaptive estimation, aggregation, copula, Kullback-Leibler divergence, log-additive exponential model, maximum entropy, nonparametric estimation, order statistics, probability density, spectral density

Remerciements

Tout d'abord, je tiens à remercier mes directeurs de thèse Cristina Butucea, Jean-François Delmas et Anne Dutfoy. Ce trio d'encadrants complémentaires m'a permis de travailler dans les meilleures conditions scientifiques et techniques. Je remercie Cristina pour l'introduction au monde des statistiques non-paramétriques, et pour ses contributions indispensables surtout à la deuxième partie de la thèse. Merci à Anne pour l'encadrement industriel au sein d'EDF. Elle a fait tous pour que cette thèse soit une réussite académique ainsi qu'une véritable contribution à la recherche appliquée menée chez EDF. Enfin je remercie Jean-François pour son soutien continu tout au long de ces trois ans. Son aptitude à comprendre et résoudre des problèmes complexes, sa rigueur pour la rédaction et sa capacité d'expliquer clairement les choses me servait d'exemple à suivre, et j'en tirais énormément d'expérience. Je les remercie également pour leur confiance de m'avoir accepté en thèse sans connaissances préalables, et leurs disponibilités pour des réunions fréquentes qui ont facilité l'avancement des travaux.

Je remercie Jean-David Fermanian et Johan Segers pour avoir accepté d'être rapporteurs de ma thèse. Leurs remarques m'ont permis d'améliorer le manuscrit. C'est le cours de Jean-David sur les copules à l'ENSAE qui a attiré mon attention sur ce domaine de la théorie de probabilités. Je suis très heureux d'avoir Agnès Sulem dans mon jury, qui m'a chaleureusement recommandé à mes futurs directeurs, ce qui a rendu cette thèse possible. Je remercie également Alexandre Tsybakov de faire partie du jury en tant qu'examinateur, ses articles étaient une grande source d'inspiration pour mes travaux statistiques.

Je suis très reconnaissant à Régis Lebrun pour les nombreux échanges sur la première partie de la thèse qui concerne les copules de statistiques d'ordres, et son aide avec le logiciel OpenTURNS. Sans lui, je n'aurais pas des résultats numériques dans ce manuscrit.

Je remercie le département Management des Risques Industriels d'EDF Recherche et Développement pour le financement de thèse. En plus du soutien financier, les formations proposées et la possibilité de travailler avec les chercheurs expérimentés m'ont permis d'approfondir mes connaissances en probabilités et statistiques appliquées dans un contexte d'ingénierie. Je tiens à remercier en particulier la communauté APSPP, où j'avais l'occasion de présenter mes avancements d'une manière régulière. Les retours donnés par les participants étaient très formateurs et donnaient des idées intéressantes pour l'ensemble de la thèse. Je remercie également l'Association Nationale de la Recherche et de la Technologie pour sa contribution au financement sous forme de la convention CIFRE.

Pour l'excellent encadrement académique, je présente mes remerciements à l'École Doctorale MSTIC de l'Université Paris-Est, et plus précisément le laboratoire CERMICS où j'ai passé la plupart de mon temps en tant que thésard. Je voudrais remercier en particulier Jean-Philippe Chancelier et Régis Monneau pour des discussions sur des questions d'optimisation, et dans un sens plus large tous mes collègues avec qui j'ai eu de la chance de travailler. Coté administration, merci à Isabelle Simunic et Sylvie Cach pour avoir rendu ma vie plus facile.

J'aimerais également remercier tous les doctorants, stagiaires, jeunes embauchés de CERMICS et d'EDF pour les pauses café étendues caractérisées par des débats intéressants et plein

de bonne humeur. Les afterworks passés ensemble étaient toujours les points culminants de la semaine. Je voudrais y rajouter tous les doctorants et chercheurs avec qui j'ai fait connaissance lors des conférences de proba et stat, et également tous mes camarades du Master de Dauphine. Je suis très heureux d'avoir pu rencontrer autant de gens brillants et sympathiques.

Les salutations ne seraient pas complètes sans mentionner mes amis de la résidence Bleuzen à Vanves. Vous étiez ma famille en France, et c'est grâce à vous que je ne me suis jamais senti seul dans ce pays. Les soirées passées ensemble, les sorties du week-end, les longues conversations resteront toujours de beaux souvenirs, et j'ai hâte de vous revoir aussi souvent que possible.

Enfin, j'adresse toute ma gratitude à mes parents et ma soeur. Sans eux, je n'aurais jamais réussi à arriver jusqu'ici. Ils m'ont beaucoup aidé dans mon parcours. Leur soutien inconditionnel m'a donné la force d'aller jusqu'au bout. Les mots ne sont pas suffisants pour exprimer tous mes sentiments, donc je dis tout simplement : merci, merci et encore merci.

Résumé substantiel

Les travaux présentés dans cette thèse portent sur la modélisation probabiliste et l'inférence statistique des vecteurs aléatoires, qui sont soumis à des contraintes déterministes. En particulier, nous nous sommes intéressés à l'étude des statistiques d'ordre, c.à.d. des vecteurs aléatoires dont les composantes sont ordonnées presque sûrement. Après une introduction générale, le manuscrit consiste en quatre papiers publiés ou soumis à des journaux scientifiques à comité de lecture :

- C. Butucea, J.-F. Delmas, A. Dutfoy and R. Fischer. Maximum entropy copula with given diagonal section. *Journal of Multivariate Analysis*, 137, 61-81, 2015. [33]
- C. Butucea, J.-F. Delmas, A. Dutfoy and R. Fischer. Maximum entropy distribution of order statistics with given marginals. *En révision à Bernoulli*. [36]
- C. Butucea, J.-F. Delmas, A. Dutfoy and R. Fischer. Optimal exponential bounds for aggregation of estimators for the Kullback-Leibler loss. *Soumis*. [38]
- C. Butucea, J.-F. Delmas, A. Dutfoy and R. Fischer. Fast adaptive estimation of log-additive exponential models in Kullback-Leibler divergence. *Soumis*. [37]

et deux papiers qui présentent des cas d'application issus des travaux au sein du département Management des Risques Industriels à EDF R&D, paraissant dans des actes des congres :

- C. Butucea, J.-F. Delmas, A. Dutfoy and R. Fischer. Modélisation de la dépendance sous contrainte déterministe. Dans *Proceedings of Congrès Lambda Mu 19 de Maîtrise des Risques et Sûreté de Fonctionnement*, 2014. [31]
- C. Butucea, J.-F. Delmas, A. Dutfoy and R. Fischer. Nonparametric estimation of distributions of order statistics with application to nuclear engineering. Dans *Safety and Reliability of Complex Engineered Systems : ESREL 2015*, 2015. [34]

La première partie de la thèse est consacrée à l'étude des statistiques d'ordre dont les distributions marginales sont fixées. Étant données ces contraintes, la loi jointe des statistiques d'ordre peut être caractérisée en précisant sa *copule* associée. Une copule est la fonction de répartition jointe d'un vecteur aléatoire $U = (U_1, \dots, U_d)$ telle que U_i est distribuée uniformément sur l'intervalle $I = [0, 1]$ pour tous $1 \leq i \leq d$. Le théorème de Sklar nous assure que la fonction de répartition jointe F_X d'un vecteur de statistiques d'ordre $X = (X_1, \dots, X_d)$ s'écrit comme la composition d'une copule et les fonctions de répartition marginales F_i de X_i :

$$F_X(x) = C(F_1(x_1), \dots, F_d(x_d)) \qquad \text{pour } x = (x_1, \dots, x_d) \in \mathbb{R}^d.$$

De plus, la copule figurant dans cette décomposition est unique si les $(F_i, 1 \leq i \leq d)$ sont continues. Dans la suite, on considère que les marginales sont continues. Parmi les lois jointes compatibles avec les contraintes, nous cherchons celle qui contient la moindre information supplémentaire par rapport aux contraintes. Soit h une densité de référence sur $\mathbb{R}$ et $h^{\otimes d}(x) = \prod_{i=1}^d h(x_i)$ pour $x = (x_1, \dots, x_d) \in \mathbb{R}^d$. On mesure la quantité d'information d'un vecteur aléatoire X avec fonction de répartition F_X par l'entropie de Shannon relative $H_h(F_X)$ définie comme :

$$H_h(F_X) = \begin{cases} -\int f_X \log\left(f_X / h^{\otimes d}\right) & \text{si } F_X \text{ est absolument continue avec densité } f_X, \\ -\infty & \text{sinon.} \end{cases}$$

Notez que $H_h(F_X) \in [-\infty, 0]$ est bien définie. Quand $h = \mathbf{1}_I$, on écrit tout simplement $H(F_X)$ au lieu de $H_h(F_X)$. On cherche à maximiser ce critère parmi les lois jointes admissibles. Par Lemme 3.1, l'entropie relative d'un vecteur aléatoire $X = (X_1, \dots, X_d)$ se décompose comme

la somme des entropies relatives marginales $H_h(F_i)$, $1 \le i \le d$ et l'entropie $H(C)$ d'un vecteur aléatoire U dont la fonction de répartition est la copule C associée à X :

$$H_h(F_X) = \sum_{i=1}^{d} H_h(F_i) + H(C).$$

Par conséquent, la maximisation de l'entropie pour des statistiques d'ordre à marginales fixées est équivalente à la maximisation de l'entropie de la copule associée.

Dans le Chapitre 2, qui correspond à [33], nous considérons le problème de trouver la copule d'entropie maximale avec une diagonale fixée. Ce problème est fondamentalement lié au problème de copule d'entropie maximale pour des statistiques d'ordre à marginales fixées, que l'on expliquera plus tard. La *diagonale* $\delta : I \to I$ d'une copule C est la fonction définie comme $\delta(t) = C(t, \dots, t)$. Si $U = (U_1, \dots, U_d)$ est un vecteur aléatoire avec fonction de répartition jointe C, alors la diagonale est la fonction de répartition de $\max_{1\le i\le d} U_i$. La diagonale porte des informations sur la dépendance des queues de la copule, et peut caractériser la fonction génératrice d'une copule Archimédienne sous une certaine condition, c.f. [77]. La solution de ce problème repose sur la théorie de maximisation d'entropie abstraite de [25].

On donne une condition nécessaire et suffisante sur la diagonale δ pour l'existence d'une copule d'entropie maximale. Notamment, une telle copule existe si et seulement si :

$$\mathcal{J}(\delta) = \int_I |\log(t - \delta(t))| \, dt < +\infty. \tag{1}$$

Cette condition est plus forte que celle de [102] pour l'existence d'une copule absolument continue avec diagonale δ, qui requiert que l'ensemble $\Sigma_\delta = \{t \in I, \delta(t) = t\}$ soit de mesure nulle par rapport à la mesure de Lebesgue. Cette dernière condition est bien assurée par (1). Quand (1) est vérifiée, on donne la formule explicite de la densité de la copule d'entropie maximale ainsi que la valeur exacte de son entropie.

D'abord, on considère le cas $\Sigma_\delta = \{0, 1\}$, c.à.d. $\delta(t) > t$ pour tous $t \in (0, 1)$. On définit les fonctions a et b comme, pour $t \in I$:

$$a(t) = \frac{d - \delta'(t)}{d} h(t)^{-1+1/d} \, \mathrm{e}^{F(t)} \quad \text{et} \quad b(t) = \frac{\delta'(t)}{d} h(t)^{-1+1/d} \, \mathrm{e}^{-(d-1)F(t)},$$

avec h et F données par :

$$h(t) = t - \delta(t), \quad F(t) = \frac{d-1}{d} \int_{\frac{1}{2}}^{t} \frac{1}{h(s)} \, ds.$$

Ces fonctions nous permettent de définir la copule $\bar{C}_\delta$, qui a pour densité $\bar{c}_\delta$ donnée par :

$$\bar{c}_\delta(x) = b(\max(x)) \prod_{x_i \ne \max(x)} a(x_i) \qquad \text{pour } x \in I^d, \tag{2}$$

avec $\max(x) = \max_{1\le i\le d} x_i$.

Dans le cas général, la continuité de la diagonale δ nous assure que $I \setminus \Sigma_\delta = \cup_{j\in J}(\alpha_j, \beta_j)$ avec J au plus dénombrable. Pour chaque $j \in J$, on écrit $\Delta_j = \beta_j - \alpha_j$, et on définit les fonctions δ^j comme :

$$\delta^j(t) = \frac{\delta(\alpha_j + t\Delta_j) - \alpha_j}{\Delta_j} \qquad \text{pour } t \in I.$$

On peut vérifier que δ^j est également la diagonale d'une copule qui satisfait $\Sigma_{\delta^j} = \{0, 1\}$. Soit $\bar{c}_{\delta^j}$ définie par (2) avec δ remplacée par δ^j. Enfin, soit C_δ la copule dont la densité c_δ est donnée par :

$$c_\delta(x) = \sum_{j\in J} \frac{1}{\Delta_j} \bar{c}_{\delta^j}\left(\frac{x - \alpha_j \mathbf{1}}{\Delta_j}\right) \mathbf{1}_{(\alpha_j,\beta_j)^d}(x) \qquad \text{pour } x \in I^d, \tag{3}$$

avec $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^d$. Quand $\Sigma_\delta = \{0, 1\}$, les fonctions c_δ et $\bar{c}_\delta$ coïncident. Le résultat principal du Chapitre 2 stipule que quand (1) est vérifiée, la copule C_δ est celle d'entropie maximale avec diagonale δ. Notons par $\mathcal{C}^\delta$ l'ensemble de tous les copules avec diagonale δ.

Théorème. *Soit δ la diagonale d'une copule.*

a) Si $\mathcal{J}(\delta) = +\infty$, alors $\max_{C\in\mathcal{C}^\delta} H(C) = -\infty$.

b) Si $\mathcal{J}(\delta) < +\infty$, alors $\max_{C\in\mathcal{C}^\delta} H(C) > -\infty$, et $C_\delta \in \mathcal{C}^\delta$, dont la densité c_δ est donnée par (3), est l'unique copule telle que : $H(C_\delta) = \max_{C\in\mathcal{C}^\delta} H(C)$. De plus, on a :

$$H(C_\delta) = -(d-1)\mathcal{J}(\delta) + \mathcal{G}(\delta),$$

où $\mathcal{G}(\delta) \in \mathbb{R}$ est donnée par :

$$\mathcal{G}(\delta) = d\log(d) + (d-1) + H(\delta) - \int_I (d-\delta')\log(d-\delta').$$

Nous illustrons la différence entre les copules issues des familles classiques et les copules d'entropie maximale avec la même diagonale à la fin de ce chapitre.

Dans le Chapitre 3, qui correspond à [36], on résout le problème initial de trouver la loi jointe d'entropie maximale pour les statistiques d'ordre à marginales $\mathbf{F} = (F_i, 1 \le i \le d)$ fixées. Nous avons déjà constaté que ce problème est équivalent à trouver la copule d'entropie maximale compatible avec les contraintes. D'après [119] les copules compatibles sont exactement celles dont le support est inclus dans un sous-ensemble de I^d qui dépend de $\mathbf{F}$. Vu qu'une contrainte sur le support est difficile à traiter avec le formalisme de [25], nous établissons une bijection entre les copules des statistiques d'ordre avec marginales $\mathbf{F}$ et un nouvel ensemble des copules contraintes. Ce dernier ensemble est celui des copules symétriques avec une multidiagonale fixée. La *multidiagonale* $\boldsymbol{\delta} = (\delta_{(1)}, \dots, \delta_{(d)})$ d'une copule C est la généralisation de la diagonale. C'est le vecteur des fonctions de répartition $\delta_{(i)}$ de $U_{(i)}$ pour $1 \le i \le d$, où $U_{(i)}$ est la i-ème plus grand composante du vecteur aléatoire U dont la fonction de répartition est la copule C. Autrement dit, $(U_{(1)}, \dots, U_{(d)})$ est le vecteur des statistiques d'ordre associée à U. Remarquons que $\delta_{(d)}$ correspond à la diagonale δ définie précédemment. La multidiagonale fixée des copules dans le nouvel ensemble est une fonction des marginales $\mathbf{F}$. La bijection préserve l'entropie à une constante additive près. Donc il faut que l'on trouve la copule symétrique d'entropie maximale avec une multidiagonale fixée. Ce problème peut être résolu d'une manière similaire au problème des copules d'entropie maximales à diagonale fixée. En fait, pour $d = 2$, ils sont équivalents.

De manière similaire à (1), la condition d'existence d'une copule d'entropie maximale avec multidiagonale $\boldsymbol{\delta}$ s'écrit comme :

$$\mathbb{J}(\boldsymbol{\delta}) = \sum_{i=2}^{d} \int_I \delta_{(i)}(dt) \left|\log\left(\delta_{(i-1)}(t) - \delta_{(i)}(t)\right)\right| < +\infty. \tag{4}$$

Pour donner la formule exacte de la solution, nous introduisons quelques notations. Notons :

$$\Psi_i^{\boldsymbol{\delta}} = \{t \in I,\, \delta_{(i-1)}(t) > \delta_i(t)\} \quad \text{pour } 2 \le i \le d.$$

L'ensemble complémentaire $(\Psi_i^{\boldsymbol{\delta}})^c$ sur I est la collection des points où $\delta_{(i-1)} = \delta_{(i)}$. On définit l'ensemble $\Sigma^{\boldsymbol{\delta}} \subset I$ comme $\Sigma^{\boldsymbol{\delta}} = \bigcup_{i=2}^d \delta_{(i)}\left((\Psi_i^{\boldsymbol{\delta}})^c\right)$. Selon [103], il existe une copule absolument continue avec multidiagonale $\boldsymbol{\delta}$ si $\Sigma^{\boldsymbol{\delta}}$ est de mesure nulle. Celle-ci est bien assurée lorsque (4) est vérifiée. Comme les ensembles $\Psi_i^{\boldsymbol{\delta}}$ sont ouverts, il existe un ensemble au plus dénombrable d'intervalles disjointes $\{(g_i^{(j)}, d_i^{(j)}),\, j \in J_i\}$ tel que $\Psi_i^{\boldsymbol{\delta}} = \bigcup_{j\in J_i}(g_i^{(j)}, d_i^{(j)})$ pour $2 \le i \le d$. Soient $m_i^{(j)} = (g_i^{(j)} + d_i^{(j)})/2$ les points milieux de ces intervalles. L'ensemble $L_{\boldsymbol{\delta}}$ est défini par :

$$L_{\boldsymbol{\delta}} = \{u = (u_1, \dots, u_d) \in I^d; (u_{(i-1)}, u_{(i)}) \subset \Psi_i^{\boldsymbol{\delta}} \text{ pour tout } 2 \le i \le d\}.$$

Considérons la copule $C_{\boldsymbol{\delta}}$ avec densité jointe $c_{\boldsymbol{\delta}}$ donnée par, pour $x = (x_1, \dots, x_d) \in I^d$:

$$c_{\boldsymbol{\delta}}(x) = \frac{1}{d!}\,\mathbf{1}_{L_{\boldsymbol{\delta}}}(x) \prod_{i=1}^{d} a_i(x_{(i)}), \tag{5}$$

où $x_{(i)}$ est la i-ème plus grande composante de x, et les fonctions a_i, $1 \leq i \leq d$, sont données par, pour $t \in I$:

$$a_i(t) = K_i'(t)\, \mathrm{e}^{K_{i+1}(t)-K_i(t)}\, \mathbf{1}_{\Psi_i^{\boldsymbol{\delta}} \cap \Psi_{i+1}^{\boldsymbol{\delta}}}(t),$$

avec pour $1 \leq i \leq d$, $t \in (g_i^{(j)}, d_i^{(j)})$:

$$K_i(t) = \int_{m_i^{(j)}}^{t} \frac{\delta'_{(i)}(s)}{\delta_{(i-1)}(s) - \delta_{(i)}(s)}\, ds$$

et les conventions $\Psi_1^{\boldsymbol{\delta}} = (0, d_1)$ avec $d_1 = \inf\{t \in I; \delta_{(1)}(t) = 1\}$, $m_1 = 0$, $\Psi_{d+1}^{\boldsymbol{\delta}} = (g_{d+1}, 1)$ avec $g_{d+1} = \sup\{t \in I; \delta_{(d)}(t) = 0\}$, $m_{d+1} = (1 + g_{d+1})/2$, $\delta_{(0)} = 1$ et $K_{d+1} = 0$.

La solution du problème est résumée dans le théorème suivant. Notons par $\mathcal{C}^{\boldsymbol{\delta}}$ l'ensemble des copules de multidiagonale $\boldsymbol{\delta}$.

Théorème. *Soit $\boldsymbol{\delta}$ la multidiagonale d'une copule.*

(a) Si $\mathbb{J}(\boldsymbol{\delta}) = +\infty$, alors $\max_{C \in \mathcal{C}^{\boldsymbol{\delta}}} H(C) = -\infty$.

(b) Si $\mathbb{J}(\boldsymbol{\delta}) < +\infty$, alors $\max_{C \in \mathcal{C}^{\boldsymbol{\delta}}} H(C) > -\infty$ et $C_{\boldsymbol{\delta}}$ dont la densité $c_{\boldsymbol{\delta}}$ est donnée par (5) est l'unique copule telle que $H(C_{\boldsymbol{\delta}}) = \max_{C \in \mathcal{C}^{\boldsymbol{\delta}}} H(C)$. De plus, on a :

$$H(C_{\boldsymbol{\delta}}) = -\mathbb{J}(\boldsymbol{\delta}) + \log(d!) + (d-1) + \sum_{i=1}^{d} H(\delta_{(i)}).$$

Puisque la solution est une copule symétrique, en appliquant la bijection sur celle-ci nous retrouvons la copule d'entropie maximale pour des statistiques d'ordre à marginales $\mathbf{F}$ fixées. Par le théorème de Sklar, on peut identifier, quand elle existe, la loi jointe d'entropie maximale pour des statistiques d'ordre à marginales $\mathbf{F}$ fixées. Une telle distribution existe si et seulement si $H_h(F_i) > -\infty$ avec X_i distribuée selon F_i pour tous $1 \leq i \leq d$, et si on a en plus :

$$\mathbb{J}(\mathbf{F}) = \sum_{i=2}^{d} \int_{\mathbb{R}} F_i(dt)\ \left|\log\left(F_{i-1}(t) - F_i(t)\right)\right| < +\infty.$$

Dans ce cas, la densité $f_{\mathbf{F}}$ de la distribution maximisant l'entropie prend la forme, pour $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$:

$$f_{\mathbf{F}}(x) = f_1(x_1) \prod_{i=2}^{d} \frac{f_i(x_i)}{F_{i-1}(x_i) - F_i(x_i)} \exp\left(- \int_{x_{i-1}}^{x_i} \frac{f_i(s)}{F_{i-1}(s) - F_i(s)}\, ds \right) \mathbf{1}_{L^{\mathbf{F}}}(x), \tag{6}$$

où f_i la densité correspondant à F_i et $L^{\mathbf{F}} \subset \mathbb{R}^d$ est l'ensemble des vecteurs ordonnés $(x_1, \ldots, x_d)$, tels que $F_{i-1}(t) > F_i(t)$ pour tous $t \in (x_{i-1}, x_i)$ et $2 \leq i \leq d$. Le résultat principal de cette partie de la thèse est le théorème suivant. Notons par $\mathcal{L}_d^{OS}(\mathbf{F})$ l'ensemble des fonctions de répartition des statistiques d'ordre avec marginales $\mathbf{F}$.

Théorème. *Soit $\mathbf{F} = (F_i, 1 \leq i \leq d)$ un vecteur de fonctions de répartition tel que $F_{i-1} \geq F_i$ pour tous $2 \leq i \leq d$ et h une densité de référence sur $\mathbb{R}$.*

(a) S'il existe $1 \leq i \leq d$ tel que $H_h(F_i) = -\infty$, ou si $\mathbb{J}(\mathbf{F}) = +\infty$, alors $\max_{F \in \mathcal{L}_d^{OS}(\mathbf{F})} H_h(F) = -\infty$.

(b) Si $H_h(F_i) > -\infty$ pour tous $1 \leq i \leq d$, et $\mathbb{J}(\mathbf{F}) < +\infty$, alors $\max_{F \in \mathcal{L}_d^{OS}(\mathbf{F})} H_h(F) > -\infty$, et la fonction de répartition jointe $F_{\mathbf{F}}$ avec densité jointe $f_{\mathbf{F}}$ définie dans (6) est la fonction de répartition unique dans $\mathcal{L}_d^{OS}(\mathbf{F})$ telle que $H_h(F_{\mathbf{F}}) = \max_{F \in \mathcal{L}_d^{OS}(\mathbf{F})} H_h(F)$. De plus, on a :

$$H_h(F_{\mathbf{F}}) = d - 1 + \sum_{i=1}^{d} H_h(F_i) - \mathbb{J}(\mathbf{F}).$$

Nous remarquons que la densité $f_{\mathbf{F}}$ a une forme produit sur $L^{\mathbf{F}}$, c.à.d. que l'on peut l'écrire, pour $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$:

$$f_{\mathbf{F}}(x) = \prod_{i=1}^{d} p_i(x_i) \mathbf{1}_{L^{\mathbf{F}}}(x), \tag{7}$$

avec des fonctions non-négatives $(p_i, 1 \leq i \leq d)$.

Le Chapitre 4, qui a donné lieu à la présentation [31], présente une application de la loi jointe d'entropie maximale des statistiques d'ordre avec des marginales fixées pour modéliser des paramètres d'entrée pour un code de calcul. Ce code simule une procédure de soudage, basé sur une méthode d'éléments finis pour un modèle thermomécanique. Il évalue les caractéristiques des fissures résiduelles qui peuvent apparaître dans le matériel pendant le soudage, ayant un impact sévère sur la durée de vie du composant soudé. Le but de cette simulation est de réaliser des études d'analyse de sensibilité sur les valeurs des paramètres d'entrée afin d'identifier ceux qui ont le plus d'impact sur la sortie du code. Les paramètres d'entrée correspondent à des caractéristiques physiques du matériel comme le module de Young, la limite d'élasticité, etc. Ces paramètres sont des fonctions monotones de la température, évalués sur une plage de température couvrant une large gamme. Alors que les valeurs des paramètres pour les basses températures sont relativement bien connues, les données sont rares pour des températures élevées, ce qui conduit à des incertitudes que l'on doit prendre en compte.

Dans ce chapitre, nous proposons d'utiliser la loi d'entropie maximale des statistiques d'ordre avec des marginales fixées pour remplacer la méthode actuelle qui consiste à imposer la valeur moyenne à chaque température et puis à ajouter une fonction d'erreur multipliée par un bruit aléatoire centré. Cette dernière approche présente plusieurs inconvénients : elle implique une hypothèse assez forte sur la forme de la courbe des paramètres, et elle peut conduire à des profils de paramètres non-monotones. Enfin, elle ne donne pas assez de flexibilité pour la modélisation des marginales individuelles. La modélisation que l'on propose résout ces problèmes : elle permet de choisir les distributions des marginales (lorsqu'elles sont stochastiquement ordonnées), elle respecte la monotonie, et des formules simples sont disponibles pour générer des réalisations de la loi obtenue. Le cas avec des marginales uniformes est discuté en détail.

Dans la deuxième partie de la thèse, on étudie le problème d'estimation non-paramétrique des densités d'entropie maximales des statistiques d'ordre obtenues dans le Chapitre 3. Selon l'équation (7), ces densités ont une forme produit sur un sous-ensemble du simplexe $S = \{x = (x_1, \ldots, x_d) \in \mathbb{R}^d, x_1 \leq \ldots, \leq x_d\}$. Cette structure spéciale suggère qu'une méthode conçue spécifiquement pour estimer ces densités jointes pourrait atteindre une vitesse de convergence univariée, évitant le fléau de la dimension qui impacte fortement la performance des méthodes d'estimation usuelles. Dans le cadre non-paramétrique, on suppose que la densité jointe appartient à un ensemble large de fonctions avec certaines propriétés de régularité, indexées par un paramètre r. Nous nous sommes particulièrement intéressés à des méthodes *adaptatives*, c.à.d. des méthodes qui n'utilisent pas de connaissances sur la régularité de la densité estimée, mais réalise toutefois la vitesse de convergence optimale pour des valeurs multiples du paramètre r. Tout au long de cette partie, on mesure la qualité de l'estimateur $\hat{f}_n$ basé sur un échantillon $\mathbb{X}^n = (X^1, \ldots, X^n)$ par la divergence de Kullback-Leibler, qui est une semi-distance entre des fonctions non-négatives f et g, donnée par :

$$D(f \| g) = \int f \log\left(\frac{f}{g}\right) - \int f + \int g.$$

Le chapitre 5, qui correspond à [38], présente une méthode d'agrégation convexe sur les logarithmes des estimateurs pour le problème de sélection de modèle en déviation. Cette méthode permet de créer un estimateur adaptatif de la densité d'entropie maximale des statistiques d'ordre. Considérons un modèle probabiliste $\mathcal{P} = \{\mathrm{P}_f;\ f \in \mathcal{F}\}$, où f est un paramètre de dimension infini qui caractérise la loi P_f. On note par $\mathbb{P}_f$ la probabilité par rapport à la distribution

P_f. Supposons que l'on possède un échantillon $\mathbb{X}^n = (X^1, \ldots, X^n)$ du modèle et des estimateurs $(f_k, 1 \leq k \leq N)$ indépendants de $\mathbb{X}^n$. Le but est de proposer un estimateur agrégé $\hat{f}_n$ de f qui vérifie, pour tous $x > 0$:

$$\mathbb{P}_f \left(D\left(f \| \hat{f}_n\right) > \min_{1 \leq k \leq N} D\left(f \| f_k\right) + R_{n,N,x} \right) \leq \mathrm{e}^{-x},$$

avec un terme résiduel $R_{n,N,x}$ le plus petit possible. On considère la classe des fonctions dont les logarithmes sont bornés par rapport à une mesure de référence h, i.e. $\mathcal{G} = \{f : \mathbb{R}^d \to \mathbb{R}^+, \| \log(f/h) \|_\infty < +\infty\}$.

D'abord, on étudie le problème d'estimation de densité de probabilité, où $\mathbb{X}^n$ est un échantillon i.i.d. de la densité f. Quand f et $(f_k, 1 \leq k \leq N)$ appartiennent à $\mathcal{G}$, on peut les écrire comme :

$$f = \mathrm{e}^{t-\psi} h \qquad \text{et} \qquad f_k = \mathrm{e}^{t_k - \psi_k} h, \tag{8}$$

où t et $(t_k, 1 \leq k \leq N)$ sont des fonctions telles que $\int th = 0$, $\int t_k h = 0$, et ψ, ψ_k sont des constantes de normalisation. L'estimateur agrégé $\hat{f}_n$ sera choisi dans l'ensemble $\{f_\lambda^D, \lambda \in \Lambda^+\}$ dont les éléments sont donnés par :

$$f_\lambda^D = \mathrm{e}^{t_\lambda - \psi_\lambda} h, \quad \text{avec} \quad t_\lambda = \sum_{k=1}^N \lambda_k t_k \quad \text{et} \quad \psi_\lambda = \log\left(\int \mathrm{e}^{t_\lambda} \, h \right), \tag{9}$$

avec :

$$\Lambda^+ = \{\lambda = (\lambda_k, 1 \leq k \leq N) \in \mathbb{R}^N, \lambda_k \geq 0 \text{ et } \sum_{1 \leq k \leq N} \lambda_k = 1\}. \tag{10}$$

Les poids λ d'agrégation sont déterminés à l'aide de l'échantillon $\mathbb{X}^n$. On pose $\hat{f}_n = f^D_{\hat{\lambda}^D_*}$, où $\hat{\lambda}^D_* \in \Lambda^+$ maximise un critère de maximum de vraisemblance pénalisé $H_n^D(\lambda)$ donné par :

$$H_n^D(\lambda) = \frac{1}{n} \sum_{j=1}^n t_\lambda(X^j) - \psi_\lambda - \frac{1}{2} \mathrm{pen}^D(\lambda), \tag{11}$$

avec la pénalité :

$$\mathrm{pen}^D(\lambda) = \sum_{k=1}^N \lambda_k \, D\left(f_\lambda^D \| f_k\right) = \sum_{k=1}^N \lambda_k \psi_k - \psi_\lambda.$$

Le théorème suivant démontre que si les densités jointes appartiennent à $\mathcal{F}^D(L) = \{f \in \mathcal{G};\ \| t \|_\infty \leq L\}$ pour $L > 0$, le terme résiduel $R_{n,N,x}$ pour l'estimateur $f^D_{\hat{\lambda}^D_*}$ est de l'ordre de $(\log(N) + x)/n$.

Théorème. *Soient $L, K > 0$. Soient $f \in \mathcal{F}^D(L)$ et $(f_k, 1 \leq k \leq N)$ des éléments de $\mathcal{F}^D(K)$ tels que $(t_k, 1 \leq k \leq N)$ sont linéairement indépendants. Soient $\mathbb{X}^n = (X^1, \ldots, X^n)$ un échantillon i.i.d. de densité f. Soit $f^D_{\hat{\lambda}^D_*}$ définie par (9) avec $\hat{\lambda}^D_* = \mathrm{argmax}_{\lambda \in \Lambda^+} H_n^D(\lambda)$. Alors pour tout $x > 0$, on a :*

$$\mathbb{P}_f \left(D\left(f \| f^D_{\hat{\lambda}^D_*}\right) - \min_{1 \leq k \leq N} D\left(f \| f_k\right) > \frac{\beta(\log(N) + x)}{n} \right) \leq \mathrm{e}^{-x},$$

avec $\beta = 2\exp(6K + 2L) + 4K/3$.

Le théorème suivant assure que le terme résiduel $R_{n,N,x} = (\log(N) + x)/n$ est en fait optimal.

Théorème. *Soient $N \geq 2$, $L > 0$. Alors il existe N densités jointes $(f_k, 1 \leq k \leq N)$, avec $f_k \in \mathcal{F}^D(L)$ telles que pour tous $n \geq 1$, $x \in \mathbb{R}^+$ qui satisfont :*

$$\frac{\log(N) + x}{n} < 3\left(1 - \mathrm{e}^{-L}\right)^2,$$

on a :

$$\inf_{\hat{f}_n} \sup_{f\in\mathcal{F}^D(L)} \mathbb{P}_f\left(D\left(f\|\hat{f}_n\right) - \min_{1\le k\le N} D\left(f\|f_k\right) \ge \frac{\beta'\left(\log(N)+x\right)}{n}\right) \ge \frac{1}{24}\,\mathrm{e}^{-x},$$

avec l'infimum pris sur tous les estimateurs $\hat{f}_n$ basés sur l'échantillon $\mathbb{X}^n = (X^1,\dots,X^n)$, et $\beta' = 2^{-17/2}/3$.

Nous considérons le même problème pour l'estimation de densité spectrale également. Dans ce cas là, l'échantillon $\mathbb{X}^n$ correspond à n observations consécutives d'un processus Gaussien stationnaire $(X^k, k\in\mathbb{Z})$ de densité spectrale f. Dans la définition de la classe $\mathcal{G}$, on choisit $h = 1/(2\pi)\mathbf{1}_{[-\pi,\pi]}$ comme densité de référence. Comme la densité spectrale est une fonction non-négative sur $[-\pi,\pi]$, non nécessairement d'intégrale 1, on écrit :

$$f = \frac{1}{2\pi}\,\mathrm{e}^{g}\,\mathbf{1}_{[-\pi,\pi]} \qquad \text{et} \qquad f_k = \frac{1}{2\pi}\,\mathrm{e}^{g_k}\,\mathbf{1}_{[-\pi,\pi]}.$$

L'estimateur agrégé $\hat{f}_n$ est choisi dans l'ensemble $\{f_\lambda^S, \lambda\in\Lambda^+\}$ basé sur des combinaisons convexes des fonctions $(g_k, 1\le k\le N)$:

$$f_\lambda^S = \frac{1}{2\pi}\,\mathrm{e}^{g_\lambda}\,\mathbf{1}_{[-\pi,\pi]} \quad \text{avec} \quad g_\lambda = \sum_{k=1}^N \lambda_k g_k. \tag{12}$$

On note l'intégrale de f_λ^S par m_λ. Les poids $\hat{\lambda}_*^S \in \Lambda^+$ d'agrégation maximisent le critère $\lambda\mapsto H_n^S(\lambda)$ donné par :

$$H_n^S(\lambda) = \int g_\lambda I_n - m_\lambda - \frac{1}{2}\,\mathrm{pen}^s(\lambda),$$

avec la pénalité $\mathrm{pen}^S(\lambda) = \sum_{k=1}^N \lambda_k\, D\left(f_\lambda^S\|f_k\right)$ et I_n définie par, pour $t\in[-\pi,\pi]$:

$$I_n(t) = \frac{\hat{\gamma}_0}{2\pi} + \frac{1}{\pi}\sum_{j=1}^{n-1}\hat{\gamma}_j\cos(jt) \quad \text{avec} \quad \hat{\gamma}_j = \frac{1}{n}\sum_{i=1}^{n-j} X^i X^{i+j},$$

où $(\hat{\gamma}_j, 0\le j\le n-1)$ sont les estimateurs empiriques des corrélations $(\gamma_j, 1\le j\le n-1)$. Remarquons que I_n est un estimateur non-paramétrique biaisé de la densité spectrale. Pour pouvoir établir un terme résiduel optimal, nous avons besoin d'une certaine régularité pour les fonctions f et $(f_k, 1\le k\le N)$. Pour une fonction périodique quelconque $\ell\in L^2([-\pi,\pi])$, considérons son développement sur la base de Fourier : $\ell(x) = \sum_{k\in\mathbb{Z}} a_k\,\mathrm{e}^{ikx}$ p.p. avec $a_k = \int_{-\pi}^{\pi}\mathrm{e}^{-ikx}\,\ell(x)\,dx$. La norme Sobolev fractionnaire $\|\ell\|_{2,r}$, $r>0$ est définie comme :

$$\|\ell\|_{2,r}^2 = \|\ell\|_{L^2(h)}^2 + \{\ell\}_{2,r}^2 \quad \text{avec} \quad \{\ell\}_{2,r}^2 = \sum_{k\in\mathbb{Z}} |k|^{2r}|a_k|^2.$$

On prend l'ensemble des fonctions paires, non-négatives dont la norme Sobolev fractionnaire est bornée pour $r>1/2$:

$$\mathcal{F}_r^S(L) = \{f\in\mathcal{G} :\, g = \log(2\pi f) \text{ vérifie } \|g\|_{2,r}\le L/\mathcal{C}_r \text{ et } g \text{ pair}\},$$

où $\mathcal{C}_r^2 = \sum_{k\in\mathbb{Z}}|k|^{-2r}$ est une constante qui dépend de r. Par l'inégalité de Cauchy-Schwarz, on a également que $\|g\|_\infty\le L$, et donc $f\in\mathcal{G}$. De plus, il existe une constante $C(r,L)$ telle que pour toutes $f\in\mathcal{F}_r^S(L)$, on a $\|2\pi f\|_{2,r}\le C(r,L)$, c.f. Lemme 5.9. Le théorème suivant assure que pour des densités spectrales appartenant à $\mathcal{F}_r^S(L)$, le terme résiduel pour l'estimateur $f_{\hat{\lambda}_*^S}^S$ est également $(\log(N)+x)/n$.

Théorème. *Soient $r>1/2$, $K,L>0$. Soient $f\in\mathcal{F}_r^S(L)$ et $(f_k, 1\le k\le N)$ des éléments de $\mathcal{F}_r^S(K)$ tels que $(g_k, 1\le k\le N)$ sont linéairement indépendants. Soit $\mathbb{X}^n = (X^1,\dots,X^n)$ un*

échantillon d'un processus Gaussien stationnaire avec densité spectrale f. Soit $f^S_{\hat\lambda^S_*}$ définie par (12) avec $\hat\lambda^S_* = \operatorname{argmax}_{\lambda\in\Lambda^+} H^S_n(\lambda)$. Alors pour tout $x>0$, on a :

$$\mathbb{P}_f\left(D\left(f\|f^S_{\hat\lambda^S_*}\right) - \min_{1\le k\le N} D\left(f\|f_k\right) > \frac{\beta(\log(N)+x)}{n} + \frac{\alpha}{n}\right) \le \mathrm{e}^{-x},$$

avec $\beta = 4(K\,\mathrm{e}^L + \mathrm{e}^{2L+3K})$ et $\alpha = 4KC(r,L)/\mathcal{C}_r$.

Le terme résiduel $(\log(N)+x)/n$ est aussi optimal selon le théorème suivant.

Théorème. *Soient $N\ge 2$, $r>1/2$, $L>0$. Il existe N densités spectrales $(f_k, 1\le k\le N)$ appartenant à $\mathcal{F}^S_r(L)$ telles que pour tous $n\ge 1$, $x\in\mathbb{R}^+$ qui satisfont :*

$$\frac{\log(N)+x}{n} < \frac{C(r,L)}{\log(N)^{2r}}$$

on a :

$$\inf_{\hat f_n}\sup_{f\in\mathcal{F}^S_r(L)} \mathbb{P}_f\left(D\left(f\|\hat f_n\right) - \min_{1\le k\le N} D\left(f\|f_k\right) \ge \frac{\beta'\left(\log(N)+x\right)}{n}\right) \ge \frac{1}{24}\,\mathrm{e}^{-x},$$

avec l'infimum pris sur tous les estimateurs $\hat f_n$ basés sur l'échantillon $\mathbb{X}^n = (X^1,\dots,X^n)$, et $\beta' = 8^{-5/2}/3$.

Dans le Chapitre 6, qui correspond à [37], on propose une méthode adaptative pour estimer des densités d'entropie maximales des statistiques d'ordre issues du Chapitre 3. On se restreint sur le cas où le support des densités est limité à l'ensemble $\triangle = \{x=(x_1,\dots,x_d)\in\mathbb{R}^d, 0\le x_1\le\ldots\le x_d\le 1\}$. D'après le Chapitre 3, la densité jointe f est sous forme produit, c.f. (7) Supposons de plus que f s'écrit comme :

$$f(x) = \exp\left(\sum_{i=1}^d \ell_i(x_i) - \mathrm{a}_0\right)\mathbf{1}_\triangle(x) \qquad \text{pour } x\in\mathbb{R}^d, \tag{13}$$

avec des fonctions ℓ_i bornées, centrées, mesurables sur I, la constante de normalisation a_0 et $\triangle = \{x=(x_1,\dots,x_d)\in\mathbb{R}^d, 0\le x_1\le\ldots\le x_d\le 1\}$. En plus de l'exemple du Chapitre 3, ce genre de densités jointes apparaissent comme les densités jointes des observations dans le modèle de troncature aléatoire, formulé dans [170]. Ce modèle a de nombreuses applications couvrant des disciplines variées comme l'astronomie [126], l'économie [98, 90], l'analyse de données de survie [118, 106, 125], etc.

Nous proposons d'estimer la densité jointe f par une famille exponentielle régulière qui prend en compte sa forme spéciale. L'idée consiste à approximer les fonctions ℓ_i en utilisant un développement limité sur une base appropriée $(\varphi_{i,k}, k\in\mathbb{N})$ pour tous $1\le i\le d$. Lorsque l'on prend $m=(m_1,\dots,m_d)$ fonctions de base pour un total de $|m| = \sum_{i=1}^d m_i$, le modèle est donné par, pour $\theta = (\theta_{i,k}; 1\le i\le d, 1\le k\le m_i)\in\mathbb{R}^{|m|}$ et $x=(x_1,\dots,x_d)\in\mathbb{R}^d$:

$$f_\theta(x) = \exp\left(\sum_{i=1}^d\sum_{k=1}^{m_i}\theta_{i,k}\varphi_{i,k}(x_i) - \psi(\theta)\right)\mathbf{1}_\triangle(x),$$

avec $\psi(\theta) = \log\left(\int_\triangle \exp\left(\sum_{i=1}^d\sum_{k=1}^{m_i}\theta_{i,k}\varphi_{i,k}(x_i)\right)dx\right)$ la constante de normalisation. Ce *modèle exponentiel log-additif* est une version multivariée du modèle présenté dans [12]. Les paramètres du modèle sont estimés par $\hat\theta_{m,n} = (\hat\theta_{m,n,i,k}; 1\le i\le d, 1\le k\le m_i)\in\mathbb{R}^{|m|}$ qui maximise la log-vraisemblance de l'échantillon $\mathbb{X}^n$:

$$\hat\theta_{m,n} = \operatorname*{argmax}_{\theta\in\mathbb{R}^{|m|}} \sum_{i=1}^d\sum_{k=1}^{m_i}\theta_{i,k}\hat\mu_{m,n,i,k} - \psi(\theta)$$

où $\hat{\mu}_{m,n,i,k} = (1/n)\sum_{j=1}^{n} \varphi_{i,k}(X_i^j)$ dénotent les moyennes empiriques. De manière équivalente, $\hat{\theta}_{m,n}$ satisfait les équations de maximum de vraisemblance :

$$\int_{\triangle} \varphi_{i,k}(x_i) f_{\hat{\theta}_{m,n}}(x)\, dx = \hat{\mu}_{m,n,i,k} \qquad \text{pour } 1 \le i \le d,\ 1 \le k \le m_i.$$

Le choix des fonctions de base $(\varphi_{i,k}, 1 \le i \le d, k \in \mathbb{N})$ sur $[0,1]$ est primordial pour obtenir des vitesses de convergence rapides. On propose une base polynomiale basée sur les polynômes de Jacobi. En particulier, les fonctions $(\varphi_{i,k}, k \in \mathbb{N})$ sont orthonormales par rapport à la mesure de Lebesgue sur $\triangle$ pour chaque $1 \le i \le d$. Néanmoins, le système complet n'est pas orthonormal, puisqu'il existe des produits scalaires non-nuls quand i varie. Les propriétés de la base sont considérées en détail dans la Section 6.6.

Le risque $D\left(f \| \hat{f}_{m,n}\right)$ entre la vraie densité f et son estimateur $\hat{f}_{m,n} = f_{\hat{\theta}_{m,n}}$ peut être décomposé en un terme de biais $D\left(f \| f_{\theta_m^*}\right)$ et un terme de variance $D\left(f_{\theta_m^*} \| \hat{f}_{m,n}\right)$, où $f_{\theta_m^*}$ est la *projection* de la densité jointe f sur le modèle exponentiel avec m fonctions de base, qui vérifie :

$$\int_{\triangle} \varphi_{i,k}(x_i) f_{\theta_m^*}(x)\, dx = \int_{\triangle} \varphi_{i,k}(x_i) f(x)\, dx \qquad \text{pour tous } 1 \le i \le d,\ 1 \le k \le m_i.$$

Pour contrôler le terme de biais, on suppose que pour tout $1 \le i \le d$ la fonction ℓ_i appartient à la classe de Sobolev $W_{r_i}^2(q_i)$ avec $r_i \in \mathbb{N}^*$ définie comme :

$$W_{r_i}^2(q_i) = \left\{ h \in L^2(q_i); h^{(r_i-1)} \text{ absolument continue et } h^{(r_i)} \in L^2(q_i) \right\},$$

où q_i est la marginale de la mesure de Lebesgue sur $\triangle$ dans la i-ème direction, c.f. (6.4). Le théorème suivant donne la vitesse de convergence du modèle exponentiel log-additif quand on fait tendre les nombres de paramètres m_i d'une façon appropriée. Nous rappelons que pour une suite de réels positifs $(a_n, n \in \mathbb{N})$, la suite de variables aléatoires $(Y_n, n \in \mathbb{N})$ est $O_{\mathbb{P}}(a_n)$ si pour tout $\varepsilon > 0$, il existe $C_\varepsilon > 0$ tel que :

$$\mathbb{P}\left(|Y_n/a_n| > C_\varepsilon\right) < \varepsilon \quad \text{pour tout } n \in \mathbb{N}.$$

Théorème. *Soit f une densité jointe de la forme (13). Supposons que les fonctions ℓ_i appartiennent à des classes de Sobolev $W_{r_i}^2(q_i)$, $r_i \in \mathbb{N}$ avec $r_i > d$ pour tout $1 \le i \le d$. Soit $\mathbb{X}^n$ un échantillon i.i.d. de f. On considère la suite $(m(n) = (m_1(n), \ldots, m_d(n)), n \in \mathbb{N}^*)$ telle que $\lim_{n\to\infty} m_i(n) = +\infty$ pour tous $1 \le i \le d$, et :*

$$\lim_{n\to\infty} |m|^{2d} \left(\sum_{i=1}^{d} m_i^{-2r_i} \right) = 0 \quad \text{et} \quad \lim_{n\to\infty} \frac{|m|^{2d+1}}{n} = 0.$$

La divergence de Kullback-Leibler $D\left(f \| \hat{f}_{m,n}\right)$ entre f et l'estimateur $\hat{f}_{m,n}$ converge en probabilité vers 0 avec la vitesse :

$$D\left(f \| \hat{f}_{m,n}\right) = O_{\mathbb{P}}\left(\sum_{i=1}^{d} m_i^{-2r_i} + \frac{|m|}{n} \right).$$

De plus, la convergence est uniforme sur la classe des fonctions $\mathcal{K}_r(L)$, donnée pour $L > 0$ par :

$$\mathcal{K}_r(L) = \left\{ f(x) = \exp\left(\sum_{i=1}^{d} \ell_i(x_i) - \mathrm{a}_0 \right) \mathbf{1}_{\triangle}(x) \text{ une densité } ; \| \ell_i \|_\infty \le L, \| (\ell_i)^{(r_i)} \|_{L^2(q_i)} \le L \right\}.$$

Autrement dit, nous avons la borne supérieure suivante pour la vitesse de convergence en probabilité :

$$\lim_{C\to\infty} \limsup_{n\to\infty} \sup_{f \in \mathcal{K}_r(L)} \mathbb{P}\left(D\left(f \| \hat{f}_{m,n}\right) \ge \left(\sum_{i=1}^{d} m_i^{-2r_i} + \frac{|m|}{n} \right) C \right) = 0.$$

Pour chaque m_i, $1 \leq i \leq d$, le choix optimal de $m_i(n) = \lfloor n^{1/(2r_i+1)} \rfloor$ rend le biais et la variance égaux, qui donne alors la vitesse de convergence $\sum_{i=1}^{d} n^{-2r_i/(2r_i+1)}$. Celle-ci est de l'ordre $n^{-2\min(r)/(2\min(r)+1)}$, qui correspond à la vitesse optimale dans le cas univarié pour des classes de Sobolev avec régularité $\min(r)$ (c.f. [12, 181]). La même vitesse peut être obtenue en choisissant le même nombre de fonctions dans chaque direction, c.à.d $m^*(n) = (v^*(n), \ldots, v^*(n))$ avec $v^*(n) = \lfloor n^{1/(2\min(r)+1)} \rfloor$.

Notons que le choix optimal des nombres $m_i(n)$ de fonctions de base fait intervenir le paramètre de régularité r. Dans la plupart des cas, on ne dispose pas d'une telle information. On fait donc appel à des méthodes qui peuvent s'adapter naturellement à la régularité inconnue de la densité sous-jacente. La méthode d'adaptation que l'on propose consiste en deux étapes. L'échantillon $\mathbb{X}^n$ est séparé en deux partie $\mathbb{X}_1^n$ et $\mathbb{X}_2^n$ de taille proportionnelle à n, pour l'usage de chacune des étapes.

D'abord, on fixe une suite $(N_n, n \in \mathbb{N}^*)$ croissante telle que $\lim_{n\to\infty} N_n = +\infty$. On note :

$$\mathcal{N}_n = \left\{ \lfloor n^{1/(2(d+j)+1)} \rfloor, 1 \leq j \leq N_n \right\} \quad \text{et} \quad \mathcal{M}_n = \left\{ m = (v, \ldots, v) \in \mathbb{R}^d, v \in \mathcal{N}_n \right\}.$$

Pour $m \in \mathcal{M}_n$, soit $\hat{f}_{m,n}$ l'estimateur dans le modèle exponentiel log-additif issu de l'échantillon $\mathbb{X}_1^n$. Les estimateurs $\mathcal{F}_n = (\hat{f}_{m,n}, m \in \mathcal{M}_n)$ correspondent aux choix optimaux pour des régularités r telles que $\min(r) \in \{d + j, 1 \leq j \leq N_n\}$.

Deuxièmement, on utilise la méthode d'agrégation convexe du Chapitre 5 pour construire l'estimateur final. On dénote $\hat{\ell}_{m,n}(x) = \sum_{i=1}^{d} \sum_{k=1}^{m_i} \hat{\theta}_{i,k} \varphi_{i,k}(x_i)$ pour $x = (x_1, \ldots, x_d) \in \triangle$ afin d'alléger la notation. Rappelons l'ensemble Λ^+ donné par (10). Pour $\lambda \in \Lambda^+$, la combinaison convexe $\hat{\ell}_\lambda$ des fonctions $\hat{\ell}_{m,n}$, $m \in \mathcal{M}_n$ et la densité jointe f_λ sont définies par :

$$\hat{\ell}_\lambda = \sum_{m \in \mathcal{M}_n} \lambda_m \hat{\ell}_{m,n} \quad \text{et} \quad f_\lambda = \exp(\hat{\ell}_\lambda - \psi_\lambda) \mathbf{1}_\triangle,$$

avec $\psi_\lambda = \log\left(\int_\triangle \exp(\hat{\ell}_\lambda) \right)$ la constante de normalisation. Les poids d'agrégation $\hat{\lambda}_n^*$ sont choisis en maximisant le critère $H_n^D(\lambda)$ donné par (11).

Le théorème suivant montre que si on choisit $N_n = o(\log(n))$ tel que $\lim_{n\to\infty} N_n = +\infty$, la série d'estimateurs $f_{\hat{\lambda}_n^*}$ converge vers f avec la vitesse optimale comme si la régularité était connue.

Théorème. *Soit f une densité jointe de la forme (13). Supposons que les fonctions ℓ_i appartiennent à des classes de Sobolev $W_{r_i}^2(q_i)$, $r_i \in \mathbb{N}$ avec $r_i > d$ pour tous $1 \leq i \leq d$. Soit $\mathbb{X}^n$ un échantillon i.i.d. de densité f. Soit $N_n = o(\log(n))$ tel que $\lim_{n\to\infty} N_n = +\infty$. La divergence de Kullback-Leibler $D\left(f \| \hat{f}_{m,n}\right)$ entre f et son estimateur $f_{\hat{\lambda}_n^*}$ converge en probabilité vers 0 avec la vitesse :*

$$D\left(f \| f_{\hat{\lambda}_n^*}\right) = O_{\mathbb{P}}\left(n^{-\frac{2\min(r)}{2\min(r)+1}} \right).$$

La série d'estimateurs $f_{\hat{\lambda}_n^*}$ permet d'obtenir la vitesse optimale uniformément sur des ensembles de densités avec une régularité croissante. Soit $\mathcal{R}_n = \{j, d+1 \leq j \leq R_n\}$, où R_n satisfait les inégalités :

$$R_n \leq N_n + d, \quad R_n \leq \left\lfloor n^{\frac{1}{2(d+N_n)+1}} \right\rfloor \quad \text{et} \quad R_n \leq \frac{\log(n)}{2\log(\log(N_n))} - \frac{1}{2}.$$

Sur l'ensemble $\mathcal{R}_n$, on a la borne supérieure suivante pour la vitesse de convergence en probabilité :

$$\lim_{C\to\infty} \limsup_{n\to\infty} \sup_{r \in (\mathcal{R}_n)^d} \sup_{f \in \mathcal{K}_r(L)} \mathbb{P}\left(D\left(f \| f_{\hat{\lambda}_n^*}\right) \geq \left(n^{-\frac{2\min(r)}{2\min(r)+1}} \right) C \right) = 0.$$

L'estimateur proposé est ainsi capable de s'adapter à la régularité inconnue de la densité sous-jacente sans perte de vitesse pour un ensemble large de paramètres de régularité.

Le Chapitre 7, qui a donné lieu à la présentation [34], présente le deuxième cas d'application concernant la modélisation des paramètres d'entrée d'un code de calcul pour la propagation des fissures dans un composant mécanique. Ce code implémente un modèle physique pour évaluer le risque de l'apparition d'une rupture brutale dans le composant sous une forte pression. Nous nous concentrons sur la modélisation jointe de deux paramètres d'entrée en particulier : la longueur et la hauteur initiale des fissures. Ces variables sont naturellement liées.

Pour ce cas d'application, nous disposons d'une base de données qui provient d'inspections régulières menées dans les centrales ainsi que d'essais contrôlés. Les données disponibles suggèrent que les dimensions vérifient la contrainte d'ordre. Les approches considérées antérieurement pour modéliser ces paramètres ne tiennent pas compte de cette observation.

Dans ce chapitre, nous proposons d'utiliser l'estimateur de maximum de vraisemblance du modèle exponentiel log-additif du Chapitre 6 pour estimer la loi jointe des dimensions des fissures. Les résultats obtenus montrent que le modèle a tendance à sous-estimer le risque de rupture par rapport aux approches précédentes. Ceci peut être dû au fait qu'une rupture est plus probable quand les deux dimensions sont grandes en même temps, alors que notre modèle accorde un poids considérable à la zone où la valeur de longueur est élevée et la valeur de hauteur est faible. Pour améliorer la performance du modèle proposé, il faudrait prendre en compte cette dépendance de queue élevée, en introduisant, par exemple, un copule de référence et en remplaçant la maximisation d'entropie par la maximisation d'entropie relative à cette copule.

List of publications

Papers in peer-reviewed journals

- C. Butucea, J.-F. Delmas, A. Dutfoy and R. Fischer. Maximum entropy copula with given diagonal section. *Journal of Multivariate Analysis*, 137, 61-81, 2015. [33]
- C. Butucea, J.-F. Delmas, A. Dutfoy and R. Fischer. Maximum entropy distribution of order statistics with given marginals. *Under revision in Bernoulli*. [36]
- C. Butucea, J.-F. Delmas, A. Dutfoy and R. Fischer. Optimal exponential bounds for aggregation of estimators for the Kullback-Leibler loss. *Submitted*. [38]
- C. Butucea, J.-F. Delmas, A. Dutfoy and R. Fischer. Fast adaptive estimation of log-additive exponential models in Kullback-Leibler divergence. *Submitted*. [37]

Papers appearing in conference proceedings

- C. Butucea, J.-F. Delmas, A. Dutfoy and R. Fischer. Copule d'entropie maximale avec des statistiques d'ordre fixées. In *Proceedings of Journées de Statistiques Rennes*, 2014. [32]
- C. Butucea, J.-F. Delmas, A. Dutfoy and R. Fischer. Modélisation de la dépendance sous contrainte déterministe. In *Proceedings of Congrès Lambda Mu 19 de Maîtrise des Risques et Sûreté de Fonctionnement*, 2014. [31]
- C. Butucea, J.-F. Delmas, A. Dutfoy and R. Fischer. Estimation rapide non-paramétrique de la densité de la distribution d'entropie maximale pour les statistiques d'ordre. In *Proceedings of Journées de Statistiques Lille*, 2015. [35]
- C. Butucea, J.-F. Delmas, A. Dutfoy and R. Fischer. Nonparametric estimation of distributions of order statistics with application to nuclear engineering. In *Safety and Reliability of Complex Engineered Systems: ESREL*, 2015. [34]

Contents

Chapter 1

Introduction

The results of this thesis concern the probabilistic and statistical modelling of multivariate random vectors in presence of some constraints. Such constraints often arise in an industrial context, and may include that:

- The support of the random vector is limited. This means that if X is a d-dimensional random vector with $d \geq 2$, the probabilistic mass must be concentrated on a subset $S \subsetneq \mathbb{R}^d$.
- The marginal distributions are fixed. In this case, only the dependence structure needs to be modelled, which can be done by using copula theory.

We shall consider such constraints as we study the probabilistic modelling of order statistics, i.e. random vectors that are almost surely ordered, with given marginal distributions. In the first part, after identifying the feasible models, our aim is to find the one that contains the least information in addition to these constraints.

The second part of the thesis is dedicated to the statistical estimation of the obtained model. We present a nonparametric approach that allows us to estimate such distributions with a fast convergence rate. The method is also adaptive to the unknown smoothness of the model.

Finally, two case studies are presented which apply these methods to various industrial problems considering probabilistic safety assessment in nuclear engineering at EDF Research and Development.

1.1 Probabilistic modelling

1.1.1 Preliminaries and basic definitions

A real-valued finite random variable X can be characterized by its cumulative distribution function or, when it exists, its probability density function. The *cumulative distribution function* (cdf for short) F_X of a real-valued finite random variable X is the measurable function from $\mathbb{R}$ to $I = [0, 1]$ defined as $F_X(t) = \mathbb{P}(X \leq t)$. The cumulative distribution function F_X is a non-decreasing càdlàg (right continuous with left limits) function such that $\lim_{t \to -\infty} F_X(x) = 0$ and $\lim_{t \to +\infty} F_X(x) = 1$. If there exists a measurable function $f_X : \mathbb{R} \to \mathbb{R}^+$ such that for all $t \in \mathbb{R}$:

$$F_X(t) = \int_{-\infty}^{t} f_X(s)\, ds,$$

then F_X is *absolutely continuous*, and f_X is the *probability density function* (pdf for short) of X. The same functions can be defined for finite random vectors as well, known as the joint cumulative distribution function and the joint probability density function. The *joint cumulative distribution function* (joint cdf for short) F_X of a finite random vector $X = (X_1, \ldots, X_d)$ taking values in $\mathbb{R}^d$ is the measurable function from $\mathbb{R}^d$ to I defined as, for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$:

$$F_X(x) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d).$$

This function is also non-decreasing and càdlàg in each variable. If there exists a measurable function $f_X : \mathbb{R}^d \to \mathbb{R}^+$ such that for all $x = (x_1, \dots, x_d) \in \mathbb{R}^d$:

$$F_X(x) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_d} f_X(y_1, \dots, y_d)\, dy_d \dots dy_1,$$

then F_X is *absolutely continuous*, and f_X is the *joint probability density function* (joint pdf for short) of X.

The (joint) cumulative distribution function completely characterizes the distribution: X and Y have the same distribution if and only if $F_X = F_Y$. For a random vector $X = (X_1, \dots, X_d)$, the i-th component X_i is a real-valued random variable which we call the *i-th marginal* of X. Its cdf $F_i = F_{X_i}$, referred to as the *i-th marginal cumulative distribution function* (marginal cdf for short) can be deduced from the joint cdf of X. The i-th marginal cdf F_i of a finite random vector $X = (X_1, \dots, X_d)$ is given by, for $t \in \mathbb{R}$:

$$F_i(t) = \lim_{s \to +\infty} F_X(x^{i,t,s}), \tag{1.1}$$

where $x^{i,t,s} = (x_1^{i,t,s}, \dots, x_d^{i,t,s}) \in \mathbb{R}^d$ is given by $x_j^{i,t,s} = t\mathbf{1}_{\{j=i\}} + s\mathbf{1}_{\{j \neq i\}}$ for $1 \leq j \leq d$.

Remark 1.1. Equation (1.1) implies that the distribution of the finite random vector $X = (X_1, \dots, X_d)$ determines the distribution of the marginals X_i for all $1 \leq i \leq d$. The inverse implication is not true: the distribution of the marginals X_i, $1 \leq i \leq d$ does not characterize completely the distribution of X.

In this thesis work, particular interest is given to random vectors which are almost surely ordered. Note that the following definition is motivated by the usual definition in statistical theory, where the order statistics are obtained by sorting the components of an underlying random vector in increasing order.

Definition 1.2. A random vector $X = (X_1, \dots, X_d)$ is a *vector of order statistics* if we have:

$$\mathbb{P}(X_1 \leq X_2 \leq \dots \leq X_d) = 1.$$

The fact that X is a vector of order statistics imposes a stochastic ordering constraint on the distribution of the marginals. For X and Y real-valued random variables, Y is *stochastically greater* than X (in the usual sense) if the cdfs F_X and F_Y verify:

$$F_X(t) \geq F_Y(t) \quad \text{for all } t \in \mathbb{R}.$$

We use the notation $F_X \geq F_Y$. The next proposition asserts that marginal cdfs of a vector of order statistics are stochastically ordered, and inversely if we have a collection of stochastically ordered random variables $(X_i, 1 \leq i \leq d)$, then there exist a vector of order statistics X with marginals having the same distributions as $(X_i, 1 \leq i \leq d)$.

Proposition 1.3. *Let $X = (X_1, \dots, X_d)$ be a vector of order statistics. Then for all $2 \leq i \leq d$, we have $F_{i-1} \geq F_i$. Conversely, let $(F_i, 1 \leq i \leq d)$ be a collection of cdfs such that for all $2 \leq i \leq d$, we have $F_{i-1} \geq F_i$. Then there exist a vector of order statistics $X = (X_1, \dots, X_d)$ such that X_i has cdf F_i for all $1 \leq i \leq d$.*

1.1.2 Dependence modelling via copulas

Remark 1.1 points out that the marginal distributions do not characterize the distribution of a random vector. In addition to the distributions of the marginals, an object describing the dependence structure between the marginals is necessary to fully determine the distribution of a random vector. This object is the so called copula function (often referred to as connecting copula), and this section is dedicated to give the definitions, properties and examples related to copulas. For a more complete overview of the topic, we refer to [105] and [132].

We start by giving the definition of a copula function.

Definition 1.4. A *copula* C is a measurable function from I^d to I obtained as the restriction of the joint cdf of a random vector $U = (U_1, \ldots, U_d)$ whose marginals are uniformly distributed on I, i.e. the marginal cdf F_i of U_i, $1 \leq i \leq d$ is given by, for $t \in \mathbb{R}$:

$$F_i(t) = \min(t, 1)\mathbf{1}_{\mathbb{R}^+}(t).$$

In the monograph of [132], a purely analytic definition of the multivariate copula function is provided (see Definition 2.10.6), which is equivalent to the previous definition. The next theorem, first appearing in [162] and referred to as Sklar's theorem, shows that the joint cdf of any random vector X can be written as the composition of a copula with the marginal cdfs, and inversely the composition of a copula with any collection of cdfs yields a function that is the joint cdf of a random vector X with marginal cdfs corresponding to the initial collection.

Theorem 1.5 (Sklar [162])**.** *Let $X = (X_1, \ldots, X_d)$ be a random vector with joint cdf F_X and marginal cdfs F_i, $1 \leq i \leq d$. Then there exists a copula C such that for all $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$:*

$$F_X(x) = C(F_1(x_1), \ldots, F_d(x_d)).$$

In addition, if F_i is continuous for all $1 \leq i \leq d$, then C is unique, and we shall write C_X the copula associated to X.

Conversely, let C be a copula and $(F_i, 1 \leq i \leq d)$ a collection of cdfs. Let us define the function $F : \mathbb{R}^d \to I$ as $F(x) = C(F_1(x_1), \ldots, F_d(x_d))$ for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Then there exists a random vector $Y = (Y_1, \ldots, Y_d)$ such that F is the joint cdf of Y, and the i-th marginal cdf of Y equals F_i for all $1 \leq i \leq d$.

Remark 1.6. Sklar's theorem implies that in order to give the distribution of a random vector $X = (X_1, \ldots, X_d)$, it is sufficient to precise the marginal distribution functions F_i, $1 \leq i \leq d$, and the copula C containing all information on the dependence of the components. This allows for separate modelling of the marginals and the dependence structure.

Example 1.7 (Independence)**.** Independence between the marginals of a random vector can be characterized in terms of the copula function if all marginal cdfs are continuous. Namely, for a random vector $X = (X_1, \ldots, X_d)$ with continuous marginal cdfs $F_i, 1 \leq i \leq d$, we have that the marginals X_i are independent if and only if the the copula C_X of X is the so-called *product copula* Π defined as, for $u = (u_1, \ldots, u_d) \in I^d$:

$$\Pi(u) = \Pi_{i=1}^d u_i.$$

Going back to Remark 1.1, two random vectors with the same marginal distributions can have very different joint cdfs depending on the connecting copula of the random vectors. There are numerous ways to construct copula functions: countless parametric families exist as well as methods to create new copulas based on existing ones, see Chapters 3 and 4 of [132]. Table 1.1 present the parametric families of two-dimensional copulas occurring in Chapter 2. To illustrate the variability of joint cdfs with the same marginal cdfs, Figure 1.1 gives an example of multiple joint pdfs of two-dimensional random vectors which have the same standard normal marginal distributions, but different copulas.

We also give the definition of the diagonal section of a copula C.

Definition 1.8. The *diagonal section* $\delta_C : I \to I$ of a copula C is given by, for $t \in I$:

$$\delta_C(t) = C(t, \ldots, t).$$

If $U = (U_1, \ldots, U_d)$ is a random variable with joint cdf the copula C, then the diagonal section is the cdf of $\max(U) = \max\{U_i, 1 \leq i \leq d\}$. Diagonal sections of d-dimensional copulas can be characterized by the following properties, see [102].

Proposition 1.9. *A function $\delta : I \to I$ is the diagonal section of a copula if and only if:*

(a) δ is a cumulative function on $[0, 1]$: $\delta(0) = 0$, $\delta(1) = 1$ and δ is non-decreasing;

Family	Parameters	$C(u_1, u_2)$
Gumbel	$\theta \in [1, +\infty)$	$\exp\left(-\left[(-\log(u_1))^\theta + (-\log(u_2))^\theta\right]^{\frac{1}{\theta}}\right)$
Marshall-Olkin	$\gamma_1, \gamma_2 \in (0, 1)$	$\min(u_1^{1-\gamma_1} u_2, u_1 u_2^{1-\gamma_2})$
Farlie-Gumbel-Morgenstern	$\theta \in [-1, 1]$	$u_1 u_2 + \theta u_1 u_2 (1 - u_1)(1 - u_2)$
Ail-Mikhail-Haq	$\theta \in [-1, 1]$	$\frac{u_1 u_2}{1-\theta(1-u_1)(1-u_2)}$
Normal	$\rho \in [-1, 1]$	$\Phi_\rho\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\right)$

Table 1.1 – Parametric families of two dimensional copulas. Φ_ρ denotes the joint cumulative distribution function of a two-dimensional normal random vector with standard normal marginals and correlation parameter $\rho \in [-1, 1]$, and Φ^{-1} denotes the quantile function of the standard normal distribution.

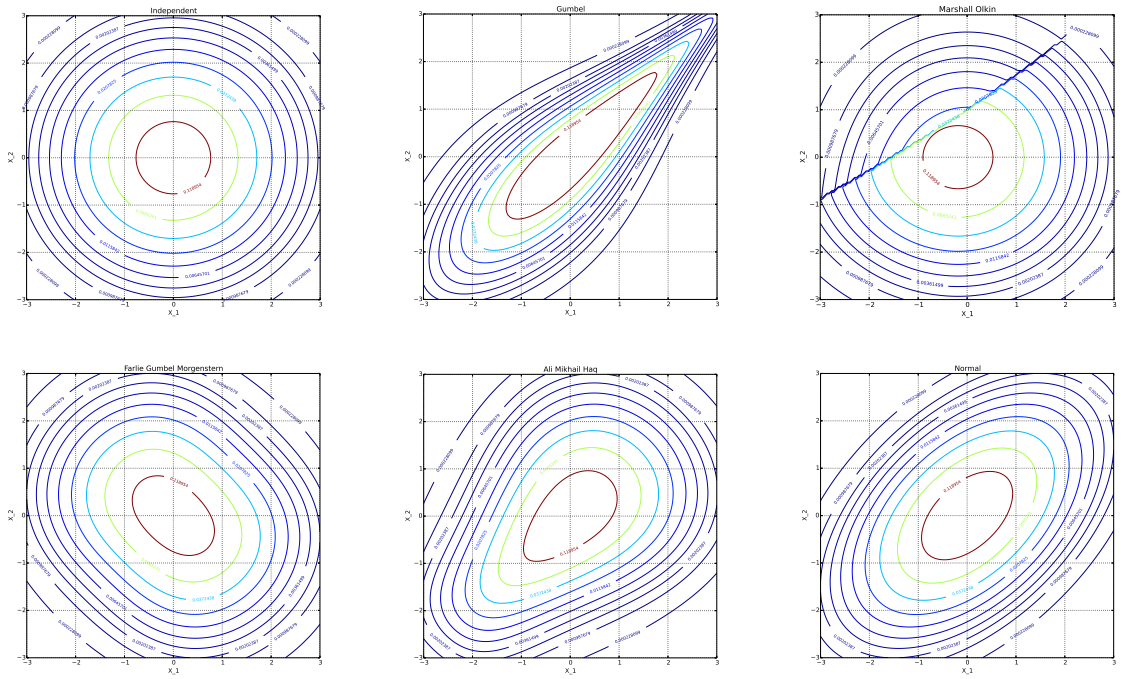


Figure 1.1 – Joint pdfs of two-dimensional random vectors with standard normal marginals and different connecting copulas.

(b) $\delta(t) \leq t$ *for* $t \in I$ *and* δ *is* d*-Lipschitz:* $|\delta(s) - \delta(t)| \leq d\,|s - t|$ *for* $s, t \in I$.

Special attention to copulas with a fixed diagonal section is given in Chapter 2.

Let us consider a two-dimensional random vector $X = (X_1, X_2)$ with continuous marginal cdfs. There exist several scalar measures which aim to quantify the stochastic dependence between X_1 and X_2. Scale-invariant measures aiming to quantify the dependence between X_1 and X_2 can be expressed in terms of the copula C_X. Two examples of widely used scale-invariant measures of concordance which only depend on the copula function are *Kendall's tau* $\tau(X)$ given by:

$$\tau(X) = 4 \int_{I^2} C_X(u_1, u_2)\, dC_X(u_1, u_2) - 1,$$

and *Spearman's rho* $\rho_S(X)$ given by:

$$\rho_S(X) = 12 \int_{I^2} u_1 u_2\, dC(u_1, u_2) - 3 = 12 \int_{I^2} C(u_1, u_2)\, du_1 du_2 - 3.$$

Note that these measures can be extended to random vectors with dimension $d > 2$, see [89]. Another measure of association of extreme values of X_1 and X_2 is given by the upper and lower tail dependence coefficient, which quantifies the dependence in the upper-right and lower left quadrant of $\mathbb{R}^2$, respectively. For $X = (X_1, X_2)$ with continuous marginal cdfs F_1 and F_2, the *upper and lower tail dependence coefficients* denoted by λ_U and λ_L respectively, are defined as:

$$\lambda_U = \lim_{t \nearrow 1} \mathbb{P}\left(X_2 > F_2^{(-1)}(t) | X_1 > F_1^{(-1)}(t)\right), \qquad \lambda_L = \lim_{t \searrow 0} \mathbb{P}\left(X_2 \leq F_2^{(-1)}(t) | X_1 \leq F_1^{(-1)}(t)\right),$$

when they exist. Notice that the upper and lower tail dependence coefficients of X can be expressed with the help of the diagonal section of its copula C_X:

$$\lambda_U = 2 - \lim_{t \nearrow 1} \frac{1 - \delta_{C_X}(t)}{1 - t}, \quad \text{and} \quad \lambda_L = \lim_{t \searrow 0} \frac{\delta_{C_X}(t)}{t}.$$

1.1.3 Measuring uncertainty by entropy

In order to be able to choose a single model out of all models which verify certain constraints, we need a decision criterion. The first attempt to give a principle comes from the works of Bernoulli and Laplace, labelled as the *principle of indifference* (or principle of insufficient reason). See Chapter I. in [143] for a discussion of this principle. It generally states that two events shall be assigned the same probability mass if we have no reason to believe that one will occur preferentially compared to the other. Aside from the lack of mathematical precision, application of this principle resulted in multiple paradoxes, especially in the case of continuous random variables. A detailed discussion of the drawbacks of this idea can be found in Chapter 4 of [113].

In the meanwhile, advances in statistical mechanics [82, 166] and information theory [161] led to the emergence of a new criteria for model selection in statistical inference based on partial knowledge, introducing a new measure of uncertainty called entropy. We give the definition of the entropy for real-valued random variables and random vectors.

Definition 1.10. Let X be a real-valued random variable (or random vector) with cdf (joint cdf) F_X. The *entropy* $H(F_X) \in \mathbb{R}$ of X (often referred to as differential entropy) is defined as:

$$H(F_X) = \begin{cases} -\int f_X \log(f_X) & \text{if } F_X \text{ is absolutely continuous with pdf (joint pdf) } f_X, \\ -\infty & \text{otherwise.} \end{cases}$$

We also say that the entropy of the real-valued random variable (or random vector) X is equal to $H(F)$.

Originating from two fundamentally different contexts, [104] shows that the two concepts are essentially the same, and advocates the acceptance of the distribution which maximizes the entropy among the admissible distributions. In the words of the author: "... making inferences on the basis of partial information we must use that probability distribution which has maximum entropy subject to whatever is known. This is the only unbiased assignment we can make; to use any other would amount to arbitrary assumption of information which by hypothesis we do not have." This is known as the *maximum entropy principal* in statistical inference. We give a few examples of maximum entropy distribution under various constraints.

Fixed interval support. The uniform distribution on the interval $[a, b]$, $a < b$ whose pdf is given by $f(t) = \mathbf{1}_{[a,b]}(t)/(b-a)$, has maximal entropy amongst real-valued random variables X such that $\mathbb{P}(a \leq X \leq b) = 1$.

Fixed expected value and positivity. The exponential distribution with parameter $\lambda \in \mathbb{R}^+$, whose pdf is given by $f(t) = \lambda \exp(-\lambda t)\mathbf{1}_{\mathbb{R}^+}(t)$, has maximal entropy amongst real-valued random variables X such that $\mathbb{P}(X > 0) = 1$ and $\mathbb{E}[X] = 1/\lambda$.

Fixed expected value and variance. The normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$, whose pdf is given by $f(t) = \exp(-(t-\mu)^2/(2\sigma^2))/\sqrt{2\pi\sigma^2}$, has maximal entropy amongst real-valued random variables X such that $\mathbb{E}[X] = \mu$ and $\mathrm{Var}\,(X) = \sigma^2$.

The differential entropy possesses some undesirable deficiencies. In particular, it can assume negative values. Furthermore, it is not invariant under parameter transformation. This led to the introduction of another measure of entropy for real-valued random variables and random vectors, which quantifies the entropy relative to a reference probability measure (and not relative to the Lebesgue measure). This measure of entropy was first introduced by Kullback and Leibler in [117]. We provide the definition for absolutely continuous real-valued random variables and random vectors.

Definition 1.11. Let X be an absolutely continuous real-valued random variable (resp. random vector) with pdf (resp. joint pdf) f_X. The *Kullback-Leibler divergence* (or relative entropy) of X with respect to a reference pdf (resp. joint pdf) q, denoted by $D\,(f_X \| q)$ is given by:

$$D\,(f_X \| q) = \int f_X \log\left(\frac{f_X}{q}\right). \tag{1.2}$$

The Kullback-Leibler divergence does not suffer from the problems of the differential entropy: it is non-negative and equals to 0 if and only if $f_X = q$ almost everywhere, and it is invariant under parameter transformation. It can also be seen as a quantity which measures the difference between (joint) pdfs, however it is not a distance on the set of (joint) pdfs, as it does not verify the triangle inequality in particular. Nevertheless, we will use the Kullback-Leibler divergence to measure the quality of estimators throughout Chapters 5, 6 and 7, as it has a natural connection to entropy.

1.1.4 Main results of the first part

In this section we summarize the main results of the first part of the thesis. In Chapter 2, which corresponds to [33], we studied maximum entropy copulas with any given diagonal section as in Definition 1.8. The problem can be phrased as follows.

> Let $\delta : I \to I$ be a function satisfying conditions (a) and (b) of Proposition 1.9. Find, when it exists, the copula with diagonal section δ that has maximal entropy.

Copulas with prescribed diagonal section received a lot of attention in the literature, see [72] for an overview on construction methods and properties of such copulas. Some recent works focus on the characterization of generators of Archimedean copulas by its diagonal section [77], singular copulas with given diagonal section [71], copulas with fixed diagonal and opposite

diagonal section [58], and an extension of the diagonal section for copulas with dimension $d \geq 3$ [73].

The solution of this problem relies on the theory outlined in [25]. The method described in this paper was used to derive the maximum entropy copula with a given Spearman's rho coefficient in [128], and more generally in a multivariate discrete setting with a given set of Spearman's rho coefficients between some components in [141]. Maximum entropy copulas with any finite number of expectation constraints are considered in [16]. Applications for maximum entropy copulas include financial modelling [61, 46, 183], tomography processing [146], hydrology [142, 97, 4], Bayesian networks [101], etc. In our work, the constraint on the diagonal section of the copula gives an infinite dimensional optimization problem as opposed to the previously cited papers where the imposed constraints are of finite dimension.

We give a necessary and sufficient condition for the existence of a maximum entropy copula with a given diagonal section. Namely, we show that there exist a unique maximum entropy copula with diagonal section δ if and only if:

$$\mathcal{J}(\delta) = \int_I |\log(t - \delta(t))| \, dt < +\infty. \tag{1.3}$$

This is a stronger condition than the condition for the existence of an absolutely continuous copula with this diagonal section, given in [102]. The condition of [102] requires that $\Sigma_\delta = \{t \in I, \delta(t) = t\}$ has zero Lebesgue measure, which is ensured whenever (1.3) holds. When (1.3) is satisfied, we give the analytic formula of the maximum entropy copula as well as the exact value of its entropy.

First, consider the case when $\Sigma_\delta = \{0, 1\}$, i.e. $\delta(t) > t$ for all $t \in (0, 1)$. Let us define the functions a and b as, for $t \in I$:

$$a(t) = \frac{d - \delta'(t)}{d} h(t)^{-1+1/d} \, \mathrm{e}^{F(t)} \quad \text{and} \quad b(t) = \frac{\delta'(t)}{d} h(t)^{-1+1/d} \, \mathrm{e}^{-(d-1)F(t)},$$

with h and F defined as:

$$h(t) = t - \delta(t), \quad F(t) = \frac{d-1}{d} \int_{\frac{1}{2}}^{t} \frac{1}{h(s)} \, ds. \tag{1.4}$$

We define the copula $\bar{C}_\delta$ with joint pdf $\bar{c}_\delta$ given by:

$$\bar{c}_\delta(x) = b(\max(x)) \prod_{x_i \neq \max(x)} a(x_i) \qquad \text{for } x \in I^d. \tag{1.5}$$

See Proposition 2.2 which verifies that $\bar{C}_\delta$ is indeed a copula with diagonal section δ.

For the general case, when Σ_δ does not necessarily euqal to $\{0, 1\}$, the continuity of δ allows us to write $I \setminus \Sigma_\delta = \cup_{j \in J} (\alpha_j, \beta_j)$ with J at most countable. For each $j \in J$, let us define $\Delta_j = \beta_j - \alpha_j$, and the function δ^j by:

$$\delta^j(t) = \frac{\delta(\alpha_j + t\Delta_j) - \alpha_j}{\Delta_j} \qquad \text{for } t \in I.$$

It is easy to verify that δ^j satisfies the conditions (a) and (b) of Proposition 1.9, therefore it is a diagonal section which verifies $\Sigma_{\delta^j} = \{0, 1\}$. Let $\bar{c}_{\delta^j}$ be defined by (1.5) with δ replaced by δ^j. Then let C_δ be the copula whose joint pdf c_δ is given by:

$$c_\delta(x) = \sum_{j \in J} \frac{1}{\Delta_j} \bar{c}_{\delta^j} \left(\frac{x - \alpha_j \mathbf{1}}{\Delta_j} \right) \mathbf{1}_{(\alpha_j, \beta_j)^d}(x) \quad \text{for } x \in I^d, \tag{1.6}$$

with $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^d$. Notice that when $\Sigma_\delta = \{0, 1\}$, then c_δ and $\bar{c}_\delta$ coincide. The main result of Chapter 2 states that when (1.3) holds, then C_δ is the maximum entropy copula with diagonal section δ. Let us denote $\mathcal{C}^\delta = \{C \text{ a copula}, \delta_C = \delta\}$, the set of all copulas whose diagonal section is δ.

Theorem 1.12. *Let δ satisfy the conditions (a) and (b) of Proposition 1.9.*

a) If $\mathcal{J}(\delta) = +\infty$ then $\max_{C\in\mathcal{C}^\delta} H(C) = -\infty$.

b) If $\mathcal{J}(\delta) < +\infty$ then $\max_{C\in\mathcal{C}^\delta} H(C) > -\infty$, and $C_\delta \in \mathcal{C}^\delta$, whose joint pdf is given by (1.6), is the unique copula such that $H(C_\delta) = \max_{C\in\mathcal{C}^\delta} H(C)$. Furthermore, we have:

$$H(C_\delta) = -(d-1)\mathcal{J}(\delta) + \mathcal{G}(\delta),$$

where $\mathcal{G}(\delta) \in \mathbb{R}$ is given by:

$$\mathcal{G}(\delta) = d\log(d) + (d-1) - \int_I \delta\log(\delta) - \int_I (d-\delta')\log(d-\delta').$$

As an illustration, we compare the maximum entropy copula to classical families of copulas, seen in Table 1.1, with the same diagonal section. Let us take, for example, the family of Farlie-Gumbel-Morgenstern copulas, given by $C(u_1,u_2) = u_1u_2 + \theta u_1u_2(1-u_1)(1-u_2)$ for $\theta \in [-1,1]$. Its diagonal section is given by, for $t \in I$:

$$\delta(t) = t^2 + \theta t^2(1-t)^2 = \theta t^4 - 2\theta t^3 + (1+\theta)t^2.$$

This diagonal section verifies $\mathcal{J}(\delta) < +\infty$ and $\Sigma_\delta = \{0,1\}$. Therefore the joint pdf c_δ of the maximum entropy copula C_δ equals to $\bar{c}_\delta$ given by (1.5). For the function F appearing in (1.4), we have:

$$F(t) = \begin{cases} \frac{1}{2}\log\left(\frac{t}{1-t}\right) + \frac{\theta}{\sqrt{4\theta-\theta^2}}\arctan\left(\frac{2\theta t-\theta}{\sqrt{4\theta-\theta^2}}\right) & \text{if } \theta \in (0,1], \\ \frac{1}{2}\log\left(\frac{t}{1-t}\right) & \text{if } \theta = 0, \\ \frac{1}{2}\log\left(\frac{t}{1-t}\right) - \frac{\theta}{\sqrt{\theta^2-4\theta}}\operatorname{arctanh}\left(\frac{2\theta t-\theta}{\sqrt{\theta^2-4\theta}}\right) & \text{if } \theta \in [-1,0). \end{cases}$$

Therefore c_δ is given by, for $\theta \in (0,1]$ and $(u_1,u_2) \in I^2$ with $u_1 \le u_2$ (by symmetry, the formula is the same for $u_2 > u_1$ with u_1, u_2 exchanged):

$$c_\delta(u_1,u_2) = \frac{(1-2\theta u_1^3 + 3\theta u_1^2 + (1+\theta)u_1)}{(1-u_1)\sqrt{\theta u_1^2 - \theta u_1 + 1}} \frac{(2\theta u_2^2 + 3\theta u_2 + (1+\theta))}{\sqrt{\theta u_2^2 - \theta u_2 + 1}} \exp\left(-\frac{\theta}{\sqrt{4\theta-\theta^2}}\left(\arctan\left(\frac{2\theta u_2-\theta}{\sqrt{4\theta-\theta^2}}\right) - \arctan\left(\frac{2\theta u_1-\theta}{\sqrt{4\theta-\theta^2}}\right)\right)\right).$$

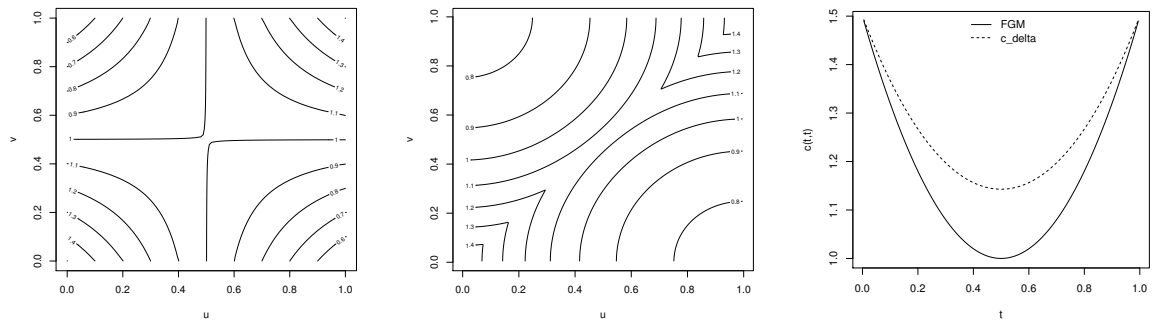
See Figure 1.2 which illustrates the difference between the joint pdfs of the Farlie-Gumbel-Morgenstern copula with parameter $\theta = 0.5$ and the maximum entropy copula C_δ with the same diagonal section, and also the difference between their diagonal cross sections $c_\delta(t,t)$, $t \in I$.

In Chapter 3, which corresponds to [36], we solve the central problem of the first part of the thesis. Let h be a reference probability density function on $\mathbb{R}$. We define $h^{\otimes d}(x) = \prod_{i=1}^d h(x_i)$ for $x = (x_1,\dots,x_d) \in \mathbb{R}^d$. We denote the relative entropy of a cdf F to $h^{\otimes d}$ by:

$$H_h(F) = \begin{cases} -\int f\log\left(f/h^{\otimes d}\right) & \text{if } F \text{ is absolutely continuous with pdf } f, \\ -\infty & \text{otherwise.} \end{cases} \tag{1.7}$$

Notice that this is minus the Kullback-Leibler divergence as in Definition 1.11. The main problem can be formulated as follows.

Let $\mathbf{F} = (F_i, 1 \le i \le d)$ be a set of continuous cdfs such that $F_{i-1} \ge F_i$ for all $2 \le i \le d$. Find, when it exists, the distribution of the vector of order statistics $X = (X_1,\dots,X_d)$, whose marginals X_i have cdf F_i for all $1 \le i \le d$, and has maximal relative entropy.



(a) FGM copula (b) Maximum entropy copula (c) Diagonal cross-section

Figure 1.2 – Isodensity lines and the diagonal cross-section of the joint pdf of the Farlie-Gumbel-Morgenstern (FGM) copula with parameter $\theta = 0.5$ and the maximum entropy copula C_δ with the same diagonal section.

Since the distributions of the marginals are fixed, by Remark 1.6 the distribution of such a vector of order statistics is determined once the connecting copula is specified. The first problem consists of identifying copulas which are compatible with the constraints. According to [119], the copula C_X of a vector of order statistics $X = (X_1, \dots, X_d)$ with fixed marginal distributions is such that the support of the random vector U with joint cdf C_X is included in a specific subset of I^d which only depends on the cdfs $\mathbf{F}$.

For the next step, notice that the relative entropy of a random vector $X = (X_1, \dots, X_d)$ with joint cdf F_X can be decomposed into sum of the relative entropy of the marginals plus the entropy the copula C_X (see Lemma 3.1):

$$H_h(F_X) = \sum_{i=1}^{d} H_h(F_i) + H(C_X).$$

Therefore X has maximal entropy if its associated copula has maximal entropy. Therefore the initial problem is equivalent to finding the maximum entropy copula compatible with the constraints.

Since constraints concerning the support of the copula is hard to take into account with this approach, we transformed the maximization problem into an equivalent one, for which the constraints could be treated with the formalism of [25]. When $d = 2$, this new set of constrained copulas are symmetric copulas with a fixed diagonal section. This diagonal section is a function of the originally fixed marginal cdfs. In the case of $d \geq 3$, we need a more general object to express the constraints of the image set of the transformation. Recall that the diagonal section of a copula C is the cdf of $\max(U)$, where U is a random vector with joint cdf C. We introduce a generalization of the diagonal section.

Definition 1.13. Let C be a d-dimensional copula and U a random vector with joint cdf C. The *multidiagonal* $\boldsymbol{\delta}_C = (\delta_{(i)}, 1 \leq i \leq d)$ of C is a vector of cdfs given by, for $1 \leq i \leq d$, $t \in I$:

$$\delta_{(i)}(t) = \mathbb{P}(U_{(i)} \leq t),$$

where $U_{(i)}$ is the i-th largest component of the random vector U, that is $(U_{(1)}, \dots, U_{(d)})$ are the order statistics of U.

The diagonal section then corresponds to $\delta_{(d)}$. This object was first considered in [103], and in [154] for individual $\delta_{(i)}$ functions. A characterization of multidiagonals is given in the following lemma from [103].

Lemma 1.14. *A vector of cdfs $\boldsymbol{\delta} = (\delta_{(1)}, \dots, \delta_{(d)})$ is the multidiagonal of a copula if and only if $\delta_{(i)}$ the following conditions hold:*

$$\delta_{(i-1)} \geq \delta_{(i)} \quad \textit{for} \quad 2 \leq i \leq d, \qquad \textit{and} \qquad \sum_{i=1}^{d} \delta_{(i)}(s) = ds \quad \textit{for} \quad 0 \leq s \leq 1.$$

The transformation of the maximum entropy problem gives a bijection between copulas of order statistics and symmetric copulas with fixed multidiagonals, see Proposition 3.6. The multidiagonal of the transformed copula is a function of the fixed marginals $\mathbf{F}$. It also preserves the entropy of the copula up the an additive constant. Therefore, the problem of maxium entropy copula of vectors of order statistics with given marginals is equivalent to finding the maximum entropy symmetric copula with a given multidiagonal. As a part of Chapter 3, we solve the maximum entropy problem of copulas with a fixed multidiagonal using similar arguments as in Chapter 2, with the help of the framework of [25].

To give the solution to the problem of maximum entropy copula with a fixed multidiagonal $\boldsymbol{\delta}$, we first give a few definitions. Let:

$$\Psi_i^{\boldsymbol{\delta}} = \{t \in I,\, \delta_{(i-1)}(t) > \delta_i(t)\} \qquad \text{for } 2 \leq i \leq d. \tag{1.8}$$

The complementary set $(\Psi_i^{\boldsymbol{\delta}})^c$ on I is the collection of the points where $\delta_{(i-1)} = \delta_{(i)}$. We define $\Sigma^{\boldsymbol{\delta}} \subset I$ as $\Sigma^{\boldsymbol{\delta}} = \bigcup_{i=2}^{d} \delta_{(i)}\left((\Psi_i^{\boldsymbol{\delta}})^c\right)$. According to [103], there exist an absolutely continuous copula with multidiagonal $\boldsymbol{\delta}$ if it verifies the conditions of Lemma 1.14 and $\Sigma^{\boldsymbol{\delta}}$ has zero Lebesgue measure. Since $\Psi_i^{\boldsymbol{\delta}}$ are open subsets of I, there exist at most countably many disjoint intervals $\{(g_i^{(j)}, d_i^{(j)}),\ j \in J_i\}$ such that $\Psi_i^{\boldsymbol{\delta}} = \bigcup_{j \in J_i} (g_i^{(j)}, d_i^{(j)})$ for $2 \leq i \leq d$. We denote by $m_i^{(j)} = (g_i^{(j)} + d_i^{(j)})/2$ the midpoint of these intervals. We define the set $L_{\boldsymbol{\delta}}$ as:

$$L_{\boldsymbol{\delta}} = \{u = (u_1, \dots, u_d) \in I^d; (u_{(i-1)}, u_{(i)}) \subset \Psi_i^{\boldsymbol{\delta}} \text{ for all } 2 \leq i \leq d\},$$

Let us define the copula $C_{\boldsymbol{\delta}}$ with joint pdf $c_{\boldsymbol{\delta}}$ on I^d as, for $x = (x_1, \dots, x_d) \in I^d$:

$$c_{\boldsymbol{\delta}}(x) = \frac{1}{d!}\, \mathbf{1}_{\boldsymbol{\delta}}(x) \prod_{i=1}^{d} a_i(x_{(i)}), \tag{1.9}$$

where $x_{(i)}$ is the i-th largest component of x, and the function a_i, $1 \leq i \leq d$, are given by, for $t \in I$:

$$a_i(t) = K_i'(t)\, \mathrm{e}^{K_{i+1}(t) - K_i(t)}\, \mathbf{1}_{\Psi_i^{\boldsymbol{\delta}} \cap \Psi_{i+1}^{\boldsymbol{\delta}}}(t),$$

with for $1 \leq i \leq d$, $t \in (g_i^{(j)}, d_i^{(j)})$:

$$K_i(t) = \int_{m_i^{(j)}}^{t} \frac{\delta_{(i)}'(s)}{\delta_{(i-1)}(s) - \delta_{(i)}(s)}\, ds$$

and the conventions $\Psi_1^{\boldsymbol{\delta}} = (0, d_1)$ with $d_1 = \inf\{t \in I; \delta_{(1)}(t) = 1\}$, $m_1 = 0$, $\Psi_{d+1}^{\boldsymbol{\delta}} = (g_{d+1}, 1)$ with $g_{d+1} = \sup\{t \in I; \delta_{(d)}(t) = 0\}$, $m_{d+1} = (1 + g_{d+1})/2$, $\delta_{(0)} = 1$ and $K_{d+1} = 0$. Similarly to (1.3), the condition for the existence of a maximum entropy copula with multidiagonal $\boldsymbol{\delta}$ is:

$$\mathbb{J}(\boldsymbol{\delta}) = \sum_{i=2}^{d} \int_I \delta_{(i)}(dt)\, \left|\log\left(\delta_{(i-1)}(t) - \delta_{(i)}(t)\right)\right| < +\infty.$$

The solution for the problem is summarized in the next theorem. Let $\mathcal{C}^{\boldsymbol{\delta}}$ denote the set of all copulas with multidiagonal $\boldsymbol{\delta}$.

Theorem 1.15. *Let $\boldsymbol{\delta}$ be a vector of cdfs verifying the conditions of Lemma 1.14.*

(a) If $\mathbb{J}(\boldsymbol{\delta}) = +\infty$ then $\max_{C \in \mathcal{C}^{\boldsymbol{\delta}}} H(C) = -\infty$.

(b) If $\mathbb{J}(\boldsymbol{\delta}) < +\infty$ *then* $\max_{C\in\mathcal{C}^{\boldsymbol{\delta}}} H(C) > -\infty$ *and* $C_{\boldsymbol{\delta}}$ *with joint pdf* $c_{\boldsymbol{\delta}}$ *given by (1.9) is the unique copula such that* $H\left(C_{\boldsymbol{\delta}}\right) = \max_{C\in\mathcal{C}^{\boldsymbol{\delta}}} H(C)$*. Furthermore, we have:*

$$H(C_{\boldsymbol{\delta}}) = -\mathbb{J}(\boldsymbol{\delta}) + \log(d!) + (d-1) + \sum_{i=1}^{d} H(\delta_{(i)}).$$

Since the solution of the problem of maximum entropy copula with a fixed multidiagonal is a symmetric copula, applying the inverse of the copula transformation on the solution provides the maximum entropy copula for a vector of order statistics with given marginal distributions. This allows us, by Sklar's theorem, to identify, when it exists, the maximum entropy distribution of a vector of order statistics with fixed marginal distributions. Such a distribution exists if and only if $H_h(F_i) > -\infty$ for all $1 \le i \le d$, and:

$$\mathbb{J}(\mathbf{F}) = \sum_{i=2}^{d} \int_{\mathbb{R}} F_i(dt)\ \left|\log\left(F_{i-1}(t) - F_i(t)\right)\right| < +\infty.$$

In this case the maximum entropy vector of order statistics is absolutely continuous with joint pdf $f_{\mathbf{F}}$ defined as, for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$:

$$f_{\mathbf{F}}(x) = f_1(x_1) \prod_{i=2}^{d} \frac{f_i(x_i)}{F_{i-1}\left(x_i\right) - F_i(x_i)} \exp\left(-\int_{x_{i-1}}^{x_i} \frac{f_i(s)}{F_{i-1}(s) - F_i(s)}\, ds\right) \mathbf{1}_{L^{\mathbf{F}}}(x), \tag{1.10}$$

where f_i is the pdf corresponding to F_i and $L^{\mathbf{F}} \subset \mathbb{R}^d$ is the set of ordered vectors $(x_1, \ldots, x_d)$, that is $x_1 \le \cdots \le x_d$, such that $F_{i-1}(t) > F_i(t)$ for all $t \in (x_{i-1}, x_i)$ and $2 \le i \le d$. The main result of this part is given by the following theorem. Let $\mathcal{L}_d^{OS}(\mathbf{F})$ denote the set of joint cdfs of vectors of order statistics with marginal cdfs $\mathbf{F}$.

Theorem 1.16. *Let* $\mathbf{F} = (F_i, 1 \le i \le d)$ *be a vector of cdfs such that* $F_{i-1} \ge F_i$ *for all* $2 \le i \le d$*.*

(a) If there exists $1 \le i \le d$ *such that* $H_h(F_i) = -\infty$*, or if* $\mathbb{J}(\mathbf{F}) = +\infty$*, then we have* $\max_{F\in\mathcal{L}_d^{OS}(\mathbf{F})} H_h(F) = -\infty$*.*

(b) If $H_h(F_i) > -\infty$ *for all* $1 \le i \le d$*, and* $\mathbb{J}(\mathbf{F}) < +\infty$*, then* $\max_{F\in\mathcal{L}_d^{OS}(\mathbf{F})} H_h(F) > -\infty$*, and the joint cdf* $F_{\mathbf{F}}$ *with joint pdf* $f_{\mathbf{F}}$ *defined in (1.10) is the unique cdf in* $\mathcal{L}_d^{OS}(\mathbf{F})$ *such that* $H_h(F_{\mathbf{F}}) = \max_{F\in\mathcal{L}_d^{OS}(\mathbf{F})} H_h(F)$*. Furthermore, we have:*

$$H_h(F_{\mathbf{F}}) = d - 1 + \sum_{i=1}^{d} H_h(F_i) - \mathbb{J}(\mathbf{F}).$$

Notice that the joint pdf $f_{\mathbf{F}}$ has a product form on $L^{\mathbf{F}}$, that is it can be written as, for a.e. $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$:

$$f_{\mathbf{F}}(x) = \prod_{i=1}^{d} p_i(x_i) \mathbf{1}_{L^{\mathbf{F}}}(x), \tag{1.11}$$

with non-negative functions $(p_i, 1 \le i \le d)$. Conversely, all joint pdfs having a product form as in (1.11) correspond to a maximum entropy distribution of order statistics for some fixed marginals. Figure 1.3 shows the joint cdf and the copula of the maximum entropy distribution of order statistics with Normal marginals with unit variance and different means.

1.1.5 Perspectives

In the first part of this thesis, we considered the probabilistic modelling of random vectors with ordering and marginal constraints. In the followings, it would be interesting to consider other type of constraints, for example for a matrix $\mathbf{A} \in \mathbb{R}^{r\times d}$ and a vector $b \in \mathbb{R}^r$, one could consider constraints of the type:

$$\mathbf{A} \cdot X \le b,$$

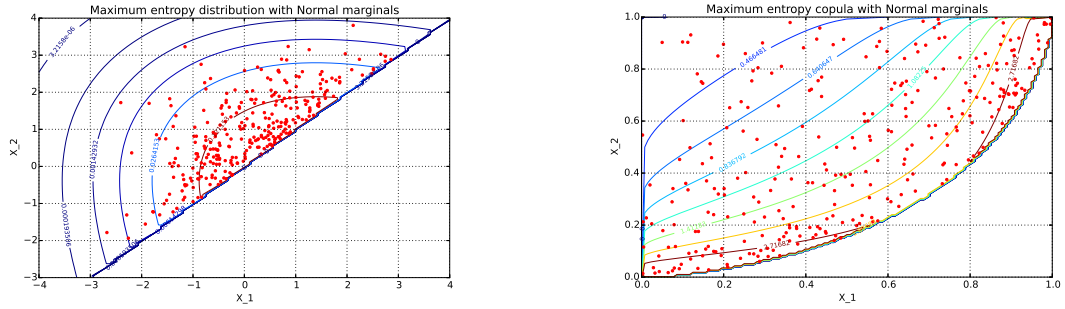


Figure 1.3 – Joint density and the copula of the maximum entropy distribution of order statistics with Normal marginals.

where $\cdot$ is the usual matrix multiplication and the inequality is understood component-wise. In particular, the ordering constraint can be written in this form with $\mathbf{A} \in \mathbb{R}^{(d-1)\times d}$ and $b \in \mathbb{R}^{d-1}$ given by, for $1 \leq i \leq d-1$ and $1 \leq j \leq d$:

$$\mathbf{A}_{ij} = \mathbf{1}_{\{i=j\}} - \mathbf{1}_{\{i=j-1\}} \quad \text{and} \quad b_i = 0.$$

More interesting and motivated by control of risks, a relaxation of the ordering constraint can be to require that the components are ordered with a certain probability, that is:

$$\mathbb{P}(X_1 \leq X_2 \leq \ldots, \leq X_d) \geq p,$$

for some $p \in (0,1)$. This would allow the random vector to deteriorate from the monotonicity constraint with probability $1-p$ introducing even more uncertainty to the problem.

For the problems of maximum entropy copulas with given diagonal section or multidiagonal, one could consider maximization of the relative entropy of the copula with respect to a reference joint pdf c_0 on I^d. Notice that maximization of the entropy for copulas is the same as maximization of the relative entropy with respect to the uniform distribution on I^d. This would allow us to incorporate further information in the modelling procedure in form of a reference pdf, to provide a more flexible framework.

1.2 Nonparametric statistical estimation

1.2.1 Nonparametric models

In many statistical problems, we consider a probabilistic model $\mathcal{P} = \{\mathrm{P}_f;\ f \in \mathcal{F}\}$ characterized by a function $f : \mathbb{R}^d \to \mathbb{R}$ for some $d \in \mathbb{N}^*$. Based on a sample $\mathbb{X}$ available from this model, the goal is to estimate f. When an explicit form for the function f is not given prior to the estimation, we have a *nonparametric* model. In the nonparametric setting, it is assumed that f belongs to a large class of functions $\mathcal{F}$ possibly with some regularity conditions. In particular, if $\mathcal{F} = \{g(x,\theta), \theta \in \Theta \subseteq \mathbb{R}^k\}$ for some $k \in \mathbb{N}^*$ with $g : \mathbb{R}^d \times \Theta \to \mathbb{R}$ a given function, we have a *parametric* model. In this case, the estimation problem is equivalent to estimating the finite-dimensional parameter θ by $\hat{\theta} = \hat{\theta}(\mathbb{X}) \in \Theta$ based on the sample. Then the parametric estimator of the function is simply given by $g(x,\hat{\theta})$. For the nonparametric model, the class $\mathcal{F}$ can not be described by a finite-dimensional parameter. Here, an estimator of f is a function $\hat{f}$ measurable with respect to the sample $\mathbb{X}$, i.e. $x \mapsto \hat{f}(x) = \hat{f}(x,\mathbb{X})$. We describe the models we consider.

Probability density estimation Let $\mathbb{X}^n = (X^1, \ldots, X^n)$ be independent, identically distributed, absolutely continuous real-valued random variables or random vectors with joint pdf f. We refer to $\mathbb{X}^n$ as an *i.i.d. sample* of size n. The problem consists of estimating f by $\hat{f}_n(x) = \hat{f}_n(x, \mathbb{X}_n)$, given that f belongs to a large class of functions $\mathcal{F}$ with some regularity conditions.

Spectral density estimation Let $(X^k, k \in \mathbb{Z})$ be a stationary sequence of centered normal random variables. *Stationarity* means that for all $n \in \mathbb{N}^*$ and all $(k_1, \dots, k_n) \in \mathbb{Z}^n$, the joint cdf of $(X^{k_1}, \dots, X^{k_n})$ equals to the joint cdf of $(X^{k_1+j}, \dots, X^{k_n+j})$ for any $j \in \mathbb{Z}$. For $j \in \mathbb{Z}$, let $\gamma_j = \mathrm{Cov}\,(X^k, X^{k+j})$ be the covariance of difference j. If $\sum_{j\in\mathbb{Z}} |\gamma_j| \leq +\infty$, then the *spectral density* associated to the sequence $(X^k, k \in \mathbb{Z})$ is the real valued, even, non-negative function $f : [-\pi, \pi] \to \mathbb{R}^+$ defined as, for $t \in [-\pi, \pi]$:

$$f(t) = \sum_{j\in\mathbb{Z}} \frac{\gamma_j}{2\pi} \mathrm{e}^{ijt} = \frac{\gamma_0}{2\pi} + \frac{1}{\pi} \sum_{j=1}^{\infty} \gamma_j \cos(jt).$$

A sample $\mathbb{X}^n = (X^k, \dots, X^{k+n-1})$ consists of the observations of n consecutive elements of the sequence $(X^k, k \in \mathbb{Z})$. The problem is to estimate the function f by $\hat{f}_n(t) = \hat{f}_n(t, \mathbb{X}^n)$, under the assumption that f belongs to a large class of functions $\mathcal{F}$ with some regularity conditions.

See [169] for other examples of nonparametric models. In the following, we concentrate on the nonparametric probability density estimation, as it is the central problem of this part of the thesis. However, in Chapter 5 we consider the spectral density estimation problem as well.

1.2.2 Nonparametric density estimation

Probably the most widely considered problem in the field of nonparametric statistics is the nonparametric probability density estimation problem defined in the previous section. A comprehensive survey on this topic is given by [100].

Convergence rates

Most of the problems concerning nonparametric density estimation aim at finding an estimator and its convergence rate which is uniform over the function class $\mathcal{F}$ with respect to a risk measure. Let $\mathbb{P}_f$ and $\mathbb{E}_f$ denote the probability and the expected value, respectively, with respect to the distribution of the sample $\mathbb{X}^n = (X^1, \dots, X^n)$ when X^j are i.i.d. with common (joint) pdf f. Let $d(f, g)$ be a semi-distance measuring the difference between the functions f and g.

Definition 1.17. A positive sequence $(\tilde{\psi}_n, n \in \mathbb{N}^*)$, which verifies $\lim_{n\to\infty} \tilde{\psi}_n = 0$ is an *upper bound for the convergence rate in expectation* of an estimator $\hat{f}_n$ over the function class $\mathcal{F}$ if there exists $C > 0$ such that:

$$\limsup_{n\to\infty} \sup_{f\in\mathcal{F}} \mathbb{E}_f[d(\hat{f}_n, f)/\tilde{\psi}_n] \leq C.$$

After establishing the rate of convergence of a certain estimator $\hat{f}_n$, the natural question which arises is: is this the best possible convergence rate we can attain for a particular problem? This led to the notion of lower bound for the convergence rate.

Definition 1.18. The sequence $(\tilde{\psi}_n, n \in \mathbb{N}^*)$ is a *lower bound for the convergence rate in expectation* if there exists $c > 0$ such that:

$$\liminf_{n\to\infty} \inf_{\hat{f}_n} \sup_{f\in\mathcal{F}} \mathbb{E}_f[d(\hat{f}_n, f)/\tilde{\psi}_n] \geq c,$$

where the infimum is taken over all estimators $\hat{f}_n$ measurable with respect to the sample $\mathbb{X}^n$.

We say that $(\tilde{\psi}_n, n \in \mathbb{N}^*)$ is the *optimal convergence rate in expectation* if it is both an upper and lower bound (it is also referred to as minimax convergence rate). In deviation, we have the following definition for convergence rates.

Definition 1.19. The sequence $(\tilde{\psi}_n, n \in \mathbb{N}^*)$ is an *upper bound for the convergence rate in deviation* of an estimator $\hat{f}_n$ over the function class $\mathcal{F}$ if:

$$\lim_{C\to\infty} \limsup_{n\to\infty} \sup_{f\in\mathcal{F}} \mathbb{P}_f(d(\hat{f}_n, f) \geq C\tilde{\psi}_n) = 0. \tag{1.12}$$

Also, $(\tilde{\psi}_n, n \in \mathbb{N}^*)$ is a *lower bound for the convergence rate in deviation* if there exists $c > 0$ such that:

$$\lim_{n\to\infty} \inf_{\hat{f}_n} \sup_{f\in\mathcal{F}} \mathbb{P}_f(d(\hat{f}_n, f) \geq c\tilde{\psi}_n) = 1. \tag{1.13}$$

An *optimal convergence rate in deviation* is both an upper and lower bound.

Nonparametric density estimation methods

Perhaps the earliest attempt to propose an estimator for the pdf without any assumption on its functional form is the *histogram*. Originally a visualization tool for datasets, the histogram appears in the literature of statistics as early as the nineteenth century. Suppose that the support of the target density f is the interval $[a, b]$. For an i.i.d. sample $\mathbb{X}^n = (X^1, \ldots, X^n)$ and a partition $[a, b] = \cup_{i=1}^m [t_n^{i-1}, t_n^i)$ where $a = t_n^0 < t_n^1 < \ldots < t_n^{m-1} < t_n^m = b$ are equally spaced with bin width h_n , the histogram estimator $\hat{f}_n^H$ is defined as, for $t \in \mathbb{R}$:

$$\hat{f}_n^H(t) = \frac{1}{nh_n} \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \mathbf{1}_{[t_n^{i-1}, t_n^i)}(X^j) \right) \mathbf{1}_{[t_n^{i-1}, t_n^i)}(t).$$

This is a maximum likelihood estimator on piecewise constant functions on the partition, see [60]. Its statistical properties when $n \to \infty$ and $nh_n \to 0$ was studied for example by [159, 79, 80].

A more sophisticated estimation method is the *kernel density estimator* $\hat{f}_n^K$ defined as, for $x \in \mathbb{R}^d$:

$$\hat{f}_n^K(x) = \frac{1}{nh_n^d} \sum_{j=1}^{n} K\left(\frac{x - X^j}{h_n}\right),$$

with $K : \mathbb{R}^d \to \mathbb{R}$ a kernel function, i.e. $\int K = 1$, and h_n the bandwidth parameter. This method was proposed by [152] and [136], and has received widespread attention. The statistical properties of the kernel estimator depends on the choice of the kernel function K [14, 39, 75, 81, 127], and the choice of the bandwidth h_n [45, 160, 94, 95, 63, 175]. Improvements proposed for the kernel density estimator include for example variable kernels [171, 28], where the bandwidth for each observation depends on the distance to its k-th nearest neighbour, or adaptive kernels [3, 165], where the bandwidth for each observation depends on a preliminary kernel estimate evaluated at the observation.

A different approach to nonparametric density estimation is the *orthogonal series density estimator* introduced by [42]. Suppose that the joint pdf f belongs to $L^2(\Omega, \nu)$ for a set $\Omega \subseteq \mathbb{R}^d$ and a reference measure ν on Ω, and that it admits the series expansion for all $x \in \Omega$:

$$f(x) = \sum_{k=0}^{\infty} \alpha_k \varphi_k(x), \tag{1.14}$$

where $\{\varphi_k, k \in \mathbb{N}\}$ is a complete orthonormal sequence of functions for $L^2(\Omega, \nu)$. The coefficients $\{\alpha_k, k \in \mathbb{N}\}$ can be calculated as $\alpha_k = \int_\Omega \varphi_k f$. This is the expected value of $\varphi_k(X)$ if the joint pdf of X is f. Therefore given a sample $\mathbb{X}^n = (X^1, \ldots, X^n)$ with joint pdf f, we can estimate $\alpha_k, k \in \mathbb{N}$ by the unbiased estimator $\hat{\alpha}_k = (1/n) \sum_{j=1}^n \varphi_k(X^j)$. The orthogonal series estimator $\hat{f}_n^O$ is obtained by taking a partial sum of r_n terms in (1.14), then plugging in it the estimators $\hat{\alpha}_k$, $1 \leq k \leq r_n$. This gives, for $x \in \Omega$:

$$\hat{f}_n^O(x) = \sum_{k=0}^{r_n} \hat{\alpha}_k \varphi_k(x).$$

The truncation point r_n plays the same role as the bandwidth h_n in the kernel density estimator. For $d = 1$, some orthonormal sequences for finitely supported pdfs include the Fourier basis for $\Omega = [0, 1]$ and ν the Lebesgue measure [172, 92], or the Legendre polynomials for $\Omega = [-1, 1]$ and ν the Lebesgue measure [93]. For infinite supports, the Hermite polynomials form on orthonormal basis for $\Omega = \mathbb{R}$ and ν the standard normal probability measure [158], and the Laguerre polynomials can be used when $\Omega = [0, +\infty)$ and ν the exponential probability measure [91].

Multiresolution analysis provides a system of orthonormal functions called the wavelet basis, which can effectively take into account discontinuities and local smoothness properties of the density function. Let us consider the one-dimensional setting. The construction of a wavelet basis relies on a scaling function or father wavelet φ and a mother wavelet ζ. The orthonormal basis consists of the functions $\{\varphi_{\ell_0,k}(x) = 2^{\ell_0/2}\varphi(2^{\ell_0}x - k), k \in \mathbb{Z}\}$ and $\{\zeta_{\ell,k}(x) = 2^{\ell/2}\zeta(2^{\ell}x - k), \ell \geq j_0, k \in \mathbb{Z}\}$. Then the *wavelet density estimator* $\hat{f}_n^W$ is given by, for $t \in \mathbb{R}$:

$$\hat{f}_n^W(t) = \sum_{k\in\mathbb{Z}} \hat{\alpha}_k \varphi_{\ell_0,k}(t) + \sum_{\ell=\ell_0}^{\infty} \sum_{k\in\mathbb{Z}} \hat{\beta}_{\ell,k} \zeta_{\ell,k}(t), \tag{1.15}$$

with $\hat{\alpha}_k = (1/n)\sum_{j=1}^n \varphi_{\ell_0,k}(X^j)$ and $\hat{\beta}_{\ell,k} = (1/n)\sum_{j=1}^n \zeta_{\ell,k}(X^j)$. Note that if the scaling function and the mother wavelet are compactly supported, then only a finite number of the coefficients $\{\hat{\alpha}_k, k \in \mathbb{Z}\}$ and $\{\hat{\beta}_{\ell,k}, \ell \geq \ell_0, k \in \mathbb{Z}\}$ are non-zero, thus $\hat{f}_n^W$ is a proper estimator. Wavelet density estimation was considered for example in [109, 110] and [173].

1.2.3 Adaptive nonparametric density estimation

Usually, the class of functions $\mathcal{F}$ considered is a class of functions with a regularity property depending on some parameter(s) r. A bound $L > 0$ is also imposed for a particular norm of the functions in $\mathcal{F}$. Therefore we use the notation $\mathcal{F} = \mathcal{F}_{r,L}$. In this case, the convergence rate of an estimator also usually depends on the regularity parameter, and shall be noted by $\tilde{\psi}_{n,r}$.

Example 1.20 (see Chapter 1.2.1. of [169])**.** Let $\mathcal{F}_{r,L}$ be the Hölder class of pdfs on $\mathbb{R}$, that is $f \in \mathcal{F}_{r,L}$ if f is a $\ell = \lfloor r \rfloor$ times differentiable pdf, and verifies $|f^{(\ell)}(t) - f^{(\ell)}(s)| \leq L\,|t - s|^{r-\ell}$ for all $t, s \in \mathbb{R}$. Let us measure the difference by the mean squared error at a fixed point $x_0 \in \mathbb{R}$, that is $d(f, g) = |f(x_0) - g(x_0)|^2$ for some $x_0 \in \mathbb{R}$. Then an upper bound for the convergence rate in expectation for the kernel density estimator with window width h_n is given by $\tilde{\psi}_{n,r} = h_n^{2r} + 1/(nh_n)$. The optimal choice of $h_n = n^{-1/(2r+1)}$ renders the two terms equal, giving the upper bound $\tilde{\psi}_{n,r} = n^{-2r/(2r+1)}$, which is also the optimal convergence rate.

As Example 1.20 shows, the construction of an optimal sequence of estimators required the knowledge of the regularity parameter r, since the definition of h_n depended on it. A more difficult problem consists of proposing an estimation procedure that does not require such extra knowledge, and can achieve the optimal convergence rate for a large set of parameters. These methods are called *adaptive estimation methods*, and it has been a key topic in the literature of nonparametric estimation for all sorts of models, leading to the emergence of multiple approaches. We give a brief overview of papers concerning adaptation methods for nonparametric density estimation.

Early papers to consider adaptive methods for nonparametric density estimation include [74] who studied data-driven linear combinations of orthogonal series estimators over Sobolev classes of periodic densities on $[0, 1]$, and [86] for Sobolev classes on $\mathbb{R}$. The so-called Lepski's method, proposed originally by [124], was also applied for the problem of point-wise adaptive estimation for Sobolev classes in [30] and adaptation for the sup-norm loss in [83]. More recently, data driven bandwidth selection methods which achieve adaptability over anisotropic Nikol'skii classes were proposed in [84, 122, 85]. Another frequently employed approach applies model selection criteria with model-complexity penalization to orthogonal series and wavelet estimators, see [11, 23, 10].

A popular adaptation method for Besov function classes for the wavelet density estimator is the wavelet thresholding procedure. This method consists of keeping only significant coefficients

in the expansion (1.15), and it was first applied to density estimation in [111] and [65]. An improvement for this technique with block thresholding rules was proposed in [96] for global error measures and [44] for local error measures such as point-wise mean squared error.

The method of aggregation of estimators, famous in machine learning and thoroughly discussed in the following section, can also be used to construct adaptive estimators. In [150], kernel density estimators are aggregated to obtain an adaptive estimator over Sobolev classes. Adaptation for the multiple index model via aggregation was considered in [155]. In Chapter 6, an aggregation method on the logarithms of density estimators, developed in Chapter 5, is used to give an adaptive estimator for densities whose logarithm belongs to a large collection of Sobolev spaces. A more detailed bibliography of aggregation methods can be found in the next section.

1.2.4 Aggregation of estimators

A multitude of nonparametric density estimation methods have been presented previously. Each method is more suited for some class of pdfs. For pdfs which belong to some parametric model, parametric estimation methods provide even faster convergence rates than nonparametric methods. The idea of proposing a unified method, which can combine the advantages of parametric models (fast convergence rate) and nonparametric methods (no fix functional form), inspired the introduction of the aggregation framework, which is attributed to [135] who formulated the problem for the nonparametric regression model. Let us give now a more detailed definition in the density estimation setup.

Let $\mathbb{X}^n = (X^1, \dots, X^n)$ denote an i.i.d. sample from a distribution with joint pdf f. Let $(f_k, 1 \leq k \leq N)$ be a collection of estimators for f, which do not depend on the sample $\mathbb{X}^n$. Consider linear combinations of these estimators: for $\mu \in \mathbb{R}^N$, let $f_\mu = \sum_{k=1}^N \mu_k f_k$. For a semi-distance $d(\cdot,\cdot)$ and a set $\mathcal{U} \subseteq \mathbb{R}^N$, the aggregation problem can be stated as follows: find an estimator $\hat{f}_n$, such that there exists a constant $C \geq 1$ for which $\hat{f}_n$ satisfies an *oracle inequality* either in expectation, that is:

$$\mathbb{E}_f \left[d(f, \hat{f}_n) \right] \leq C \min_{\mu \in \mathcal{U}} d(f, f_\mu) + R_{n,N}, \tag{1.16}$$

or in deviation, i.e. for all $\varepsilon > 0$:

$$\mathbb{P}_f \left(d(f, \hat{f}_n) > C \min_{\mu \in \mathcal{U}} d(f, f_\mu) + R_{n,N,\varepsilon} \right) \leq \varepsilon, \tag{1.17}$$

for some small remainder terms $R_{n,N}$, $R_{n,N,\varepsilon}$ independent of f and $(f_k, 1 \leq k \leq N)$ belonging to a certain class of functions. When $C = 1$, we say that the oracle inequality is *sharp*. According to the choice of the set $\mathcal{U}$, three main problems are considered in the literature.

Model selection aggregation $\mathcal{U} = \{e_k, 1 \leq k \leq N\}$, where $e_k \in \mathbb{R}^N$ denotes the unit vector in the k-th direction. This means that the aggregate estimator has to mimic the performance of the best estimator among $(f_k, 1 \leq k \leq N)$.

Convex aggregation $\mathcal{U}$ is a convex subset of $\mathbb{R}^N$, usually chosen to be the simplex:

$$\Lambda^+ = \{\mu = (\mu_k, 1 \leq k \leq N) \in \mathbb{R}^N, \mu_k \geq 0 \text{ and } \sum_{1 \leq k \leq N} \mu_k = 1\}. \tag{1.18}$$

This means that the aggregate estimator has to mimic the performance of the best convex combination of the estimators $(f_k, 1 \leq k \leq N)$.

Linear aggregation $\mathcal{U} = \mathbb{R}^N$. This means that the aggregate estimator has to mimic the performance of the best linear combination of the estimators $(f_k, 1 \leq k \leq N)$.

Notice that these problems are increasing in difficulty, since the minimum in (1.16) and (1.17) are taken over increasing sets. Hence the order of the remainder terms $R_{n,N}$ and $R_{n,N,\varepsilon}$ is specific to each problem and also increases with the difficulty. For the semi-distance $d(\cdot,\cdot)$, most papers in the literature consider the L^p distance with $1 \leq p \leq +\infty$, the Kullback-Leibler divergence or the Hellinger distance. Optimality of the remainder term is defined similarly to minimax convergence rates, see [167].

Definition 1.21. The term $R_{n,N}$ is the *optimal rate of aggregation in expectation* for functions in a class $\mathcal{F}$ if:

- there exists an aggregate estimator $\hat{f}_n$ and a constant $C > 0$ such that for all $(f_k, 1 \leq k \leq N)$, $n \in \mathbb{N}^*$:
$$\sup_{f \in \mathcal{F}} \left(\mathbb{E}_f \left[d(f, \hat{f}_n) \right] - \min_{\mu \in \mathcal{U}} d(f, f_\mu) \right) \leq C R_{n,N},$$
- there exist N functions $f_k, 1 \leq k \leq N$ in $\mathcal{F}$ and a constant $c > 0$ such that for all $n \in \mathbb{N}^*$:
$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} \left(\mathbb{E}_f \left[d(f, \hat{f}_n) \right] - \min_{\mu \in \mathcal{U}} d(f, f_\mu) \right) \geq c R_{n,N}.$$

Definition 1.22. The term $(R_{n,N,x}, n \in \mathbb{N}^*)$ is the *optimal rate of aggregation in deviation* for functions in a class $\mathcal{F}$ if for all x in some interval $(\underline{x}, \bar{x})$:

- there exists an aggregate estimator $\hat{f}_n$ and a constant $C > 0$ such that for all $(f_k, 1 \leq k \leq N)$, $n \in \mathbb{N}^*$:
$$\sup_{f \in \mathcal{F}} \mathbb{P}_f \left(d(f, \hat{f}_n) - \min_{\mu \in \mathcal{U}} d(f, f_\mu) > C R_{n,N,x} \right) \leq x,$$
- there exist N functions $f_k, 1 \leq k \leq N$ in $\mathcal{F}$ and a constant $c > 0$ such that for all $n \in \mathbb{N}^*$:
$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{P}_f \left(d(f, \hat{f}_n) - \min_{\mu \in \mathcal{U}} d(f, f_\mu) > c R_{n,N,x} \right) \geq x.$$

A lot of results on aggregation concern the nonparametric regression model with random design, which can be formulated as follows. Let $\mathbb{X}^n = ((X^1, Y^1), \ldots, (X^n, Y^n))$ be a sample of independent two-dimensional random vectors, where $Y^i, 1 \leq i \leq n$ is given by:

$$Y^i = f(X^i) + \xi^i,$$

with $f : \mathbb{R} \to \mathbb{R}$ unknown and ξ^i, $1 \leq i \leq n$ integrable real valued random variables such that $\mathbb{E}[\xi^i] = 0$. The problem is to estimate the function f by $\hat{f}_n(t) = \hat{f}_n(t, \mathbb{X}^n)$, given the assumption that f belongs to a large class of functions $\mathcal{F}$ with some regularity conditions. For the nonparametric regression model with random design, model selection aggregation was considered in [178, 174] for the L^2 distance in expectation. The procedure considered in [178], called *progressive mixture method*, is suboptimal in deviation according to [8], which proposes an alternative which is optimal both in expectation and deviation. Optimality in deviation can be achieved with restricted empirical risk minimization, see [121]. The problem of convex aggregation is addressed in [107] and [180] for large values of N, while [168] also considers linear aggregation. A universal method which achieves near optimal remainder terms in expectation for all three problems was proposed by [29]. Extension of these results for nonparametric regression with fixed design can be found in [53, 54, 50] for model selection aggregation, [52] for convex aggregation. Rates of aggregation both in expectation and deviation with respect to the Kullback-Leibler divergence for all three problems was studied in [149].

Results for model selection aggregation in density estimation were first given in [41, 179] in expectation of the Kullback-Leibler divergence. The results were shown to be optimal in [120]. A generalization of the progressive mixture method of [41, 179] is given in [108]. Model selection in deviation for L^2 distance is addressed in [17]. The problem of convex and linear aggregation for densities is considered in [150] for expectation with L^2 distance.

To our knowledge, the only paper considering aggregation of spectral density estimators is [43], where linear aggregation of lag window estimators for the L^2 distance in expectation is studied. The method was validated by a simulation study as well.

1.2.5 Main results of the second part

In this section we present the results obtained in the second part of the thesis. In Chapter 5, which corresponds to [38], we consider the problem of model-selection aggregation (that is $\mathcal{U} = \{e_k, 1 \leq k \leq N\}$) in deviation for the Kullback-Leibler divergence, defined by (1.2), with exponential bounds. Let us state the problem in a general setup.

Let $\mathbb{X}^n = (X^1, \dots, X^n)$ be a sample from the probabilistic model $\mathcal{P} = \{\mathrm{P}_f;\ f \in \mathcal{F}\}$. Let $(f_k, 1 \leq k \leq N)$ be a set of estimators independent of $\mathbb{X}^n$. Find an estimator $\hat{f}_n$ of f such that for all $x > 0$:

$$\mathbb{P}_f\left(D\left(f\|\hat{f}_n\right) > \min_{1\leq k\leq N} D\left(f\|f_k\right) + R_{n,N,x}\right) \leq \mathrm{e}^{-x},$$

with a remainder term $R_{n,N,x}$ that is optimal.

The considered class contains functions whose logarithm is bounded with respect to a reference pdf h. Let us denote:

$$\mathcal{G} = \{f : \mathbb{R}^d \to \mathbb{R}^+ \text{ measurable},\ \|\log(f/h)\|_\infty < +\infty\}$$

We first consider the density estimation problem, where f corresponds to the joint pdf from which an i.i.d. sample $\mathbb{X}^n = (X^1, \dots, X^n)$ is available. When the joint pdfs f and $(f_k, 1 \leq k \leq N)$ belong to $\mathcal{G}$, they have the following representation:

$$f = \mathrm{e}^{t-\psi}\, h \qquad \text{and} \qquad f_k = \mathrm{e}^{t_k-\psi_k}\, h, \tag{1.19}$$

where t, t_k are functions such that $\int th = 0$, $\int t_k h = 0$, and ψ, ψ_k are normalizing constants. The estimator $\hat{f}_n$ will be chosen from the family $\{f_\lambda^D, \lambda \in \Lambda^+\}$ with Λ^+ as in (1.18), whose elements are given by:

$$f_\lambda^D = \mathrm{e}^{t_\lambda-\psi_\lambda}\, h \quad \text{with} \quad t_\lambda = \sum_{k=1}^N \lambda_k t_k \quad \text{and} \quad \psi_\lambda = \log\left(\int \mathrm{e}^{t_\lambda}\, h\right). \tag{1.20}$$

Therefore the estimator $\hat{f}_n$ is based on a convex combination of the functions $(t_k, 1 \leq k \leq N)$ (rather than a convex combination of $(f_k, 1 \leq k \leq N)$), where the aggregation weights λ are determined using the sample $\mathbb{X}^n$. We set $\hat{f}_n = f^D_{\hat{\lambda}^D_*}$, where $\hat{\lambda}^D_* \in \Lambda^+$ maximizes a penalized maximum likelihood criterion, i.e. $\hat{\lambda}^D_* = \operatorname{argmax}_{\lambda\in\Lambda^+} H^D_n(\lambda)$ where $H^D_n(\lambda)$ is given by:

$$H^D_n(\lambda) = \frac{1}{n}\sum_{j=1}^n t_\lambda(X^j) - \psi_\lambda - \frac{1}{2}\,\mathrm{pen}^D(\lambda), \tag{1.21}$$

with penalty term:

$$\mathrm{pen}^D(\lambda) = \sum_{k=1}^N \lambda_k\, D\left(f^D_\lambda\|f_k\right) = \sum_{k=1}^N \lambda_k\psi_k - \psi_\lambda.$$

The following theorems show that for joint pdfs belonging to the set $\mathcal{F}^D(L) = \{f \in \mathcal{G};\ \|t\|_\infty \leq L\}$ for some $L > 0$, the estimator $f^D_{\hat{\lambda}^D_*}$ achieves the rate of aggregation $R_{n,N,x}$ given by $(\log(N) + x)/n$.

Theorem 1.23. *Let $L, K > 0$. Let $f \in \mathcal{F}^D(L)$ and $(f_k, 1 \leq k \leq N)$ be elements of $\mathcal{F}^D(K)$ such that $(t_k, 1 \leq k \leq N)$ are linearly independent. Let $\mathbb{X}^n = (X^1, \dots, X^n)$ be an i.i.d. sample from the pdf f. Let $f^D_{\hat{\lambda}^D_*}$ be given by (1.20) with $\hat{\lambda}^D_* = \operatorname{argmax}_{\lambda\in\Lambda^+} H^D_n(\lambda)$. Then for any $x > 0$ we have:*

$$\mathbb{P}_f\left(D\left(f\|f^D_{\hat{\lambda}^D_*}\right) - \min_{1\leq k\leq N} D\left(f\|f_k\right) > \frac{\beta(\log(N) + x)}{n}\right) \leq \mathrm{e}^{-x},$$

with $\beta = 2\exp(6K + 2L) + 4K/3$.

The next theorem shows that $R_{n,N,x} = (\log(N) + x)/n$ is indeed optimal.

Theorem 1.24. *Let $N \geq 2$, $L > 0$. Then there exist N pdfs $(f_k, 1 \leq k \leq N)$, with $f_k \in \mathcal{F}^D(L)$ such that for all $n \geq 1$, $x \in \mathbb{R}^+$ satisfying:*

$$\frac{\log(N) + x}{n} < 3\left(1 - \mathrm{e}^{-L}\right)^2,$$

we have:

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}^D(L)} \mathbb{P}_f\left(D\left(f \| \hat{f}_n\right) - \min_{1 \leq k \leq N} D\left(f \| f_k\right) \geq \frac{\beta'\left(\log(N) + x\right)}{n}\right) \geq \frac{1}{24}\,\mathrm{e}^{-x},$$

with the infimum taken over all estimators $\hat{f}_n$ based on the sample $\mathbb{X}^n = (X^1, \ldots, X^n)$, and $\beta' = 2^{-17/2}/3$.

We consider the same problem for spectral density estimation as well. In this case, the sample $\mathbb{X}^n$ corresponds to n consecutive observations from the stationary Gaussian sequence $(X^k, k \in \mathbb{Z})$ with spectral density f. The quality of a non-negative estimator $\hat{f}_n$ is measured by the generalized Kullback-Leibler divergence $D\left(f \| \hat{f}_n\right)$ defined as:

$$D\left(f \| \hat{f}_n\right) = \int f \log\left(\frac{f}{\hat{f}_n}\right) - \int f + \int \hat{f}_n.$$

Notice that this definition coincides with (1.2) when f and $\hat{f}_n$ are pdfs. To define the function class $\mathcal{G}$, we choose $h = 1/(2\pi)\mathbf{1}_{[-\pi,\pi]}$ as reference pdf. We give a slightly different representation of f and the estimators $(f_k, 1 \leq k \leq N)$ than (1.19), which is necessary since these functions do not necessary have unit integrals. Let:

$$f = \frac{1}{2\pi}\,\mathrm{e}^{g}\,\mathbf{1}_{[-\pi,\pi]} \qquad \text{and} \qquad f_k = \frac{1}{2\pi}\,\mathrm{e}^{g_k}\,\mathbf{1}_{[-\pi,\pi]}.$$

We choose our estimator $\hat{f}_n$ amongst the function $\{f_\lambda^S, \lambda \in \Lambda^+\}$ based on the convex combinations of the functions $(g_k, 1 \leq k \leq N)$:

$$f_\lambda^S = \frac{1}{2\pi}\,\mathrm{e}^{g_\lambda}\,\mathbf{1}_{[-\pi,\pi]} \quad \text{with} \quad g_\lambda = \sum_{k=1}^{N} \lambda_k g_k. \tag{1.22}$$

We denote the integral of f_λ^S by m_λ. The aggregation weights $\hat{\lambda}_*^S \in \Lambda^+$ maximizes the function $\lambda \mapsto H_n^S(\lambda)$ given by:

$$H_n^S(\lambda) = \int g_\lambda I_n - m_\lambda - \frac{1}{2}\,\mathrm{pen}^S(\lambda),$$

with penalty term $\mathrm{pen}^S(\lambda) = \sum_{k=1}^N \lambda_k\, D\left(f_\lambda^S \| f_k\right)$ and I_n defined as, for $t \in [-\pi, \pi]$:

$$I_n(t) = \frac{\hat{\gamma}_0}{2\pi} + \frac{1}{\pi}\sum_{j=1}^{n-1} \hat{\gamma}_j \cos(jt) \quad \text{with} \quad \hat{\gamma}_j = \frac{1}{n}\sum_{i=1}^{n-j} X^i X^{i+j},$$

where $(\hat{\gamma}_j, 0 \leq j \leq n-1)$ correspond to empirical estimates of the correlations $(\gamma_j, 1 \leq j \leq n-1)$. Notice that the function I_n is a biased nonparametric estimator of the spectral density. In order to give the optimal remainder term for this problem, we have to assume some regularity on the functions f and $(f_k, 1 \leq k \leq N)$. For a function $\ell \in L^2([-\pi, \pi])$ periodic on $[-\pi, \pi]$, let us take its Fourier series expansion: $\ell(x) = \sum_{k \in \mathbb{Z}} a_k\,\mathrm{e}^{ikx}$ a.e. with $a_k = \int_{-\pi}^{\pi} \mathrm{e}^{-ikx}\,\ell(x)\,dx$. Define the fractional Sobolev norm $\|\ell\|_{2,r}$, $r > 0$ as:

$$\|\ell\|_{2,r}^2 = \|\ell\|_{L^2(h)}^2 + \{\ell\}_{2,r}^2 \quad \text{with} \quad \{\ell\}_{2,r}^2 = \sum_{k \in \mathbb{Z}} |k|^{2r} |a_k|^2.$$

We consider non-negative, even functions whose logarithms have bounded fractional Sobolev norms for some $r > 1/2$:

$$\mathcal{F}_r^S(L) = \{f \in \mathcal{G} : \ g = \log(2\pi f) \text{ verifies } \|g\|_{2,r} \leq L/\mathcal{C}_r \text{ and } g \text{ even}\},$$

where $\mathcal{C}_r^2 = \sum_{k\in\mathbb{Z}} |k|^{-2r}$ is a constant depending on r. By the Cauchy-Schwarz inequality, we also have that $\|g\|_\infty \leq L$. There also exist a constant $C(r, L)$ such that for all $f \in \mathcal{F}_r^S(L)$, we have $\|2\pi f\|_{2,r} \leq C(r, L)$, see Lemma 5.9. The following theorems show that for spectral densities belonging to $\mathcal{F}_r^S(L)$, the estimator $f^S_{\hat{\lambda}_*^S}$ also achieves the rate of aggregation $(\log(N) + x)/n$.

Theorem 1.25. *Let $r > 1/2$, $K, L > 0$. Let $f \in \mathcal{F}_r^S(L)$ and $(f_k, 1 \leq k \leq N)$ be elements of $\mathcal{F}_r^S(K)$ such that $(g_k, 1 \leq k \leq N)$ are linearly independent. Let $\mathbb{X}^n = (X^1, \ldots, X^n)$ be a sample of a stationary centered Gaussian sequence with spectral density f. Let $f^S_{\hat{\lambda}_*^S}$ be given by (1.22) with $\hat{\lambda}_*^S = \operatorname{argmax}_{\lambda\in\Lambda^+} H_n^S(\lambda)$. Then for any $x > 0$, we have:*

$$\mathbb{P}_f\left(D\left(f\|f^S_{\hat{\lambda}_*^S}\right) - \min_{1\leq k\leq N} D\left(f\|f_k\right) > \frac{\beta(\log(N) + x)}{n} + \frac{\alpha}{n}\right) \leq \mathrm{e}^{-x},$$

with $\beta = 4(K\,\mathrm{e}^L + \mathrm{e}^{2L+3K})$ and $\alpha = 4KC(r, L)/\mathcal{C}_r$.

The rate $(\log(N) + x)/n$ is also optimal according to the next theorem.

Theorem 1.26. *Let $N \geq 2$, $r > 1/2$, $L > 0$. There exist N spectral densities $(f_k, 1 \leq k \leq N)$ belonging to $\mathcal{F}_r^S(L)$ such that for all $n \geq 1$, $x \in \mathbb{R}^+$ satisfying:*

$$\frac{\log(N) + x}{n} < \frac{C(r, L)}{\log(N)^{2r}}$$

we have:

$$\inf_{\hat{f}_n} \sup_{f\in\mathcal{F}_r^S(L)} \mathbb{P}_f\left(D\left(f\|\hat{f}_n\right) - \min_{1\leq k\leq N} D\left(f\|f_k\right) \geq \frac{\beta'\left(\log(N) + x\right)}{n}\right) \geq \frac{1}{24}\,\mathrm{e}^{-x},$$

with the infimum taken over all estimators $\hat{f}_n$ based on the sample sequence $\mathbb{X}^n = (X^1, \ldots, X^n)$, and $\beta' = 8^{-5/2}/3$.

The aggregation method proposed in this section will be used to obtain an adaptive non-parametric density estimator for maximum entropy distributions of vectors of order statistics in the following.

Chapter 6, which corresponds to [37], is devoted to the study of the problem of nonparametric estimation of maximum entropy distributions of vectors of order statistics, and can be given as follows.

Let f be the joint pdf of a d-dimensional random vector with $d \geq 2$, given by:

$$f(x) = \exp\left(\sum_{i=1}^d \ell_i(x_i) - \mathrm{a}_0\right)\mathbf{1}_\triangle(x) \qquad \text{for } x \in \mathbb{R}^d, \tag{1.23}$$

with ℓ_i bounded, centered, measurable functions on I, a_0 the normalizing constant and $\triangle = \{x = (x_1, \ldots, x_d) \in \mathbb{R}^d, 0 \leq x_1 \leq \ldots \leq x_d \leq 1\}$. Given an i.i.d. sample $\mathbb{X}^n = (X^1, \ldots, X^n)$ of size n from the pdf f, the task is to estimate nonparametrically f with a convergence rate that corresponds to the optimal rate of convergence in deviation for the univariate density estimation problem.

By (1.11), joint pdfs of maximum entropy distributions of order statistics with given marginals are of the form (1.23) if the marginals are supported on I. In addition to this example, densities of the form (1.23) appear as joint pdfs of observations in the random truncation model. The random truncation model, which was first formulated in [170], appears in various contexts ranging from astronomy [126], economics [98, 90] to survival data analysis [118, 106, 125]. See [115] for an overview of possible applications. Some recent theoretical works cover estimation of the probability of holes for the Lynden-Bell estimator [163], estimation of the mode of the density of interest [18], a Berry-Esseen type bound for the kernel density estimation under random truncation [6], etc.

For $d = 2$, let (Z_1, Z_2) be a pair of independent random variables on I such that Z_i has pdf p_i for $i \in \{1, 2\}$. In the random truncation model, we suppose that we can only observe realizations of (Z_1, Z_2) if $Z_1 \leq Z_2$. Let $\bar{Z} = (\bar{Z}_1, \bar{Z}_2)$ denote the random vector distributed as (Z_1, Z_2) conditionally on $Z_1 \leq Z_2$. Then the joint pdf f of $\bar{Z}$ is given by, for $x = (x_1, x_2) \in I^2$:

$$f(x) = \frac{1}{\alpha} p_1(x_1) p_2(x_2) \mathbf{1}_{\triangle}(x),$$

with $\alpha = \int_{I^2} p_1(x_1) p_2(x_2) \mathbf{1}_{\triangle}(x)\, dx$. The joint pdf can be written in the form of (1.23) with ℓ_i defined as $\ell_i = \log(p_i) - \int_I \log(p_i)$ for $i \in \{1, 2\}$ (when p_i is uniformly bounded from 0 and $+\infty$, and $\log(p_i)$ is integrable).

We propose to estimate joint pdfs of the form (1.23) by a regular exponential family which takes into consideration its special structure. The idea consists of approximating the function ℓ_i by an orthogonal series expansion on a suitable basis $(\varphi_{i,k}, k \in \mathbb{N})$ for all $1 \leq i \leq d$. When we take $m = (m_1, \ldots, m_d)$ basis functions for a total of $|m| = \sum_{i=1}^d m_i$, the model takes the form, for $\theta = (\theta_{i,k}; 1 \leq i \leq d, 1 \leq k \leq m_i) \in \mathbb{R}^{|m|}$ and $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$:

$$f_\theta(x) = \exp\left(\sum_{i=1}^d \sum_{k=1}^{m_i} \theta_{i,k} \varphi_{i,k}(x_i) - \psi(\theta) \right) \mathbf{1}_{\triangle}(x),$$

with $\psi(\theta) = \log\left(\int_{\triangle} \exp\left(\sum_{i=1}^d \sum_{k=1}^{m_i} \theta_{i,k} \varphi_{i,k}(x_i) \right) dx \right)$ a normalizing constant. We will refer to this model as the *log-additive exponential model*. We estimate the parameters of the model by $\hat{\theta}_{m,n} = (\hat{\theta}_{m,n,i,k}; 1 \leq i \leq d, 1 \leq k \leq m_i) \in \mathbb{R}^{|m|}$ which maximizes the log-likelihood based on the sample $\mathbb{X}^n$:

$$\hat{\theta}_{m,n} = \operatorname*{argmax}_{\theta \in \mathbb{R}^{|m|}} \sum_{i=1}^d \sum_{k=1}^{m_i} \theta_{i,k} \hat{\mu}_{m,n,i,k} - \psi(\theta)$$

with $\hat{\mu}_{m,n,i,k} = (1/n) \sum_{j=1}^n \varphi_{i,k}(X_i^j)$ the empirical means. Equivalently, the parameter estimate $\hat{\theta}_{m,n}$ satisfies the maximum likelihood equations:

$$\int_{\triangle} \varphi_{i,k}(x_i) f_{\hat{\theta}_{m,n}}(x)\, dx = \hat{\mu}_{m,n,i,k} \qquad \text{for } 1 \leq i \leq d,\ 1 \leq k \leq m_i.$$

The choice of the functions $(\varphi_{i,k}, 1 \leq i \leq d, k \in \mathbb{N})$ is crucial to obtain a fast convergence rate. We propose a polynomial basis consisting of Jacobi polynomials (transformed to suit the domain $\triangle$). In particular, the system of functions $(\varphi_{i,k}, k \in \mathbb{N})$ is orthonormal with respect to the Lebesgue measure on $\triangle$ for each $1 \leq i \leq d$. However, the complete system is not orthonormal, as some of the scalar products are non-zero. For detailed properties of the basis, see Section 6.6.

We measure the risk between the true joint pdf f and its estimator $\hat{f}_{m,n} = f_{\hat{\theta}_{m,n}}$ by the Kullback-Leibler divergence $D\left(f \| \hat{f}_{m,n}\right)$. We show that $D\left(f \| \hat{f}_{m,n}\right)$ can be separated into a bias term $D\left(f \| f_{\theta_m^*}\right)$ and a variance term $D\left(f_{\theta_m^*} \| \hat{f}_{m,n}\right)$, where $f_{\theta_m^*}$ is the so called *information projection* of the joint pdf f onto the exponential model with m basis functions, verifying:

$$\int_{\triangle} \varphi_{i,k}(x_i) f_{\theta_m^*}(x)\, dx = \int_{\triangle} \varphi_{i,k}(x_i) f(x)\, dx \qquad \text{for } 1 \leq i \leq d,\ 1 \leq k \leq m_i.$$

To control the bias term $D\left(f\|f_{\theta_m^*}\right)$, we suppose for all $1 \leq i \leq d$ that the function ℓ_i belongs to the Sobolev space $W_{r_i}^2(q_i)$, $r_i \in \mathbb{N}^*$, defined as:

$$W_{r_i}^2(q_i) = \left\{h \in L^2(q_i); h^{(r_i-1)} \text{ is absolutely continuous and } h^{(r_i)} \in L^2(q_i)\right\},$$

where q_i is the i-th marginal of the Lebesgue measure on $\triangle$. The following theorem gives the convergence rate of the maximum likelihood estimator for the log-additive exponential model when the number of parameters m_i grows with n in an appropriate manner. We recall that for a positive sequence $(a_n, n \in \mathbb{N})$, the notation $O_{\mathbb{P}}(a_n)$ of stochastic boundedness for a sequence of random variables $(Y_n, n \in \mathbb{N})$ means that for every $\varepsilon > 0$, there exists $C_\varepsilon > 0$ such that:

$$\mathbb{P}\left(|Y_n/a_n| > C_\varepsilon\right) < \varepsilon \quad \text{for all } n \in \mathbb{N}.$$

Theorem 1.27. *Let f be a joint pdf of the form (1.23). Assume the functions ℓ_i belong to the Sobolev space $W_{r_i}^2(q_i)$, $r_i \in \mathbb{N}$ with $r_i > d$ for all $1 \leq i \leq d$. Let $\mathbb{X}^n$ be an i.i.d. sample from f. We consider a sequence $(m(n) = (m_1(n), \ldots, m_d(n)), n \in \mathbb{N}^*)$ such that $\lim_{n\to\infty} m_i(n) = +\infty$ for all $1 \leq i \leq d$, and which satisfies:*

$$\lim_{n\to\infty} |m|^{2d} \left(\sum_{i=1}^d m_i^{-2r_i}\right) = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{|m|^{2d+1}}{n} = 0.$$

The Kullback-Leibler distance $D\left(f\|\hat{f}_{m,n}\right)$ between f and the maximum likelihood estimator $\hat{f}_{m,n}$ converges in probability to 0 with the convergence rate:

$$D\left(f\|\hat{f}_{m,n}\right) = O_{\mathbb{P}}\left(\sum_{i=1}^d m_i^{-2r_i} + \frac{|m|}{n}\right).$$

Furthermore, the convergence is uniform over the class of functions $\mathcal{K}_r(L)$ given by for a regularity parameter $r = (r_1, \ldots, r_d) \in (\mathbb{N}^*)^d$ and $L > 0$:

$$\mathcal{K}_r(L) = \left\{f(x) = \exp\left(\sum_{i=1}^d \ell_i(x_i) - \mathrm{a}_0\right)\mathbf{1}_{\triangle}(x) \text{ a joint pdf}; \|\,\ell_i\,\|_\infty \leq L, \|\,(\ell_i)^{(r_i)}\,\|_{L^2(q_i)} \leq L\right\}.$$

That is, we have the following upper bound for the convergence rate in deviation (as in (1.12) of Definition 1.19):

$$\lim_{C\to\infty} \limsup_{n\to\infty} \sup_{f\in\mathcal{K}_r(L)} \mathbb{P}\left(D\left(f\|\hat{f}_{m,n}\right) \geq \left(\sum_{i=1}^d m_i^{-2r_i} + \frac{|m|}{n}\right) C\right) = 0.$$

We remark that lower bounds corresponding to (1.13) of Definition 1.19 are not available in the literature for this setup. For each m_i, $1 \leq i \leq d$, the choice of $m_i(n) = \lfloor n^{1/(2r_i+1)} \rfloor$ balance out the bias and variance term giving the convergence rate $\psi_{n,r} = \sum_{i=1}^d n^{-2r_i/(2r_i+1)}$, which is the same order as $n^{-2\min(r)/(2\min(r)+1)}$, corresponding to the optimal convergence rate in the univariate case over Sobolev spaces with regularity parameter $\min(r)$ (see [12, 181]). The same rate can be obtained by choosing the same number of functions in each direction: $m^*(n) = (v^*(n), \ldots, v^*(n))$ with $v^*(n) = \lfloor n^{1/(2\min(r)+1)} \rfloor$.

Notice that similarly to the example of Section 1.2.3, the optimal choice of the number of basis functions m depended on the knowledge of the regularity parameter r. When such knowledge is not available, we propose an adaptive estimation method which achieves the convergence rate $\psi_{n,r} = n^{-2\min(r)/(2\min(r)+1)}$ for a large set of regularity parameters r. The adaptive method consists of two steps: first we estimate the log-additive exponential model for multiple choices of the number of basis functions m, which correspond to optimal choices for different regularity parameters, then in a second step we utilize the aggregation method of Chapter 5 to create a final estimator which automatically achieves the convergence rate $n^{-2\min(r)/(2\min(r)+1)}$ even if r is unknown. We split the sample $\mathbb{X}^n$ into two parts $\mathbb{X}_1^n$ and $\mathbb{X}_2^n$ of size proportional to n, to use for each step.

Estimation step Let $(N_n, n \in \mathbb{N}^*)$ be a sequence of non-decreasing positive integers such that $\lim_{n\to\infty} N_n = +\infty$. We denote:

$$\mathcal{N}_n = \left\{ \lfloor n^{1/(2(d+j)+1)} \rfloor, 1 \leq j \leq N_n \right\} \quad \text{and} \quad \mathcal{M}_n = \left\{ m = (v, \ldots, v) \in \mathbb{R}^d, v \in \mathcal{N}_n \right\}.$$

For $m \in \mathcal{M}_n$ let $\hat{f}_{m,n}$ be the maximum likelihood estimator for the log-additive exponential model based on the first sample $\mathbb{X}_1^n$. Notice that the estimators $\mathcal{F}_n = (\hat{f}_{m,n}, m \in \mathcal{M}_n)$ correspond to the optimal choices for regularity parameters r for which $\min(r) \in \{d + j, 1 \leq j \leq N_n\}$.

Aggregation step Let us write $\hat{\ell}_{m,n}(x) = \sum_{i=1}^d \sum_{k=1}^{m_i} \hat{\theta}_{i,k} \varphi_{i,k}(x_i)$ for $x = (x_1, \ldots, x_d) \in \triangle$ to ease notation. We define the convex combination $\hat{\ell}_\lambda$ of the functions $\hat{\ell}_{m,n}$, $m \in \mathcal{M}_n$ and the joint pdf f_λ as, for $\lambda \in \Lambda^+$:

$$\hat{\ell}_\lambda = \sum_{m \in \mathcal{M}_n} \lambda_m \hat{\ell}_{m,n} \quad \text{and} \quad f_\lambda = \exp(\hat{\ell}_\lambda - \psi_\lambda) \mathbf{1}_\triangle,$$

with $\psi_\lambda = \log\left(\int_\triangle \exp(\hat{\ell}_\lambda) \right)$. We choose the aggregation weights $\hat{\lambda}_n^*$ by maximizing $H_n^D(\lambda)$ given by (1.21) with $\mathbb{X}^n$ replaced by $\mathbb{X}_2^n$.

Notice that in the set $\mathcal{M}_n$ we only considered vectors with the same number of basis functions in each direction. Since we have already seen that optimal rate can be achieved by choosing such vectors, it is unnecessary to include vectors in $\mathcal{M}_n$ which correspond to anisotropic functions classes (that is, where the regularity parameters $r_1, \ldots, r_d$ are not the same in each direction).

We remark also that the aggregation step requires that the functions $\hat{\ell}_{m,n}$, $m \in \mathcal{N}_n$ be uniformly bounded. During the proof of our main result, we show that they are uniformly bounded with high probability, which is sufficient for the convergence rate in deviation.

The next theorem asserts that if we choose $N_n = o(\log(n))$ such that $\lim_{n\to\infty} N_n = +\infty$, the series of convex aggregate estimators $f_{\hat{\lambda}_n^*}$ converge to f with the optimal convergence rate, i.e. as if the smoothness was known.

Theorem 1.28. *Let f be a joint pdf of the form (1.23). Assume the functions ℓ_i belongs to the Sobolev space $W_{r_i}^2(q_i)$, $r_i \in \mathbb{N}$ with $r_i > d$ for all $1 \leq i \leq d$. Let $\mathbb{X}^n$ be an i.i.d. sample from f. Let $N_n = o(\log(n))$ such that $\lim_{n\to\infty} N_n = +\infty$. The Kullback-Leibler distance $D\left(f \| \hat{f}_{m,n}\right)$ between f and the convex aggregate estimator $f_{\hat{\lambda}_n^*}$ converges in probability to 0 with the convergence rate:*

$$D\left(f \| f_{\hat{\lambda}_n^*}\right) = O_{\mathbb{P}}\left(n^{-\frac{2\min(r)}{2\min(r)+1}} \right).$$

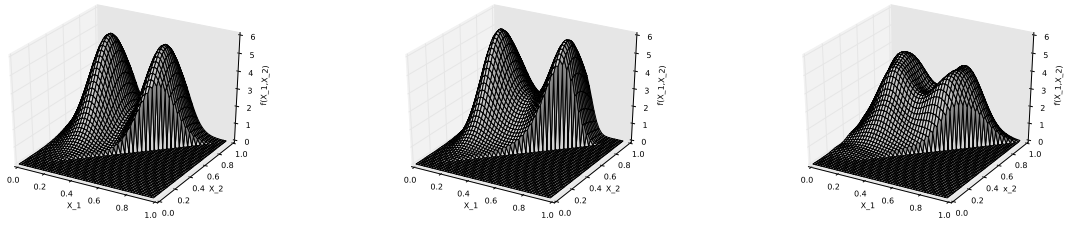
The sequence of convex aggregate estimators $f_{\hat{\lambda}_n^*}$ achieves uniform convergence over sets of densities with increasing regularity. Let $\mathcal{R}_n = \{j, d+1 \leq j \leq R_n\}$, where R_n satisfies the three inequalities:

$$R_n \leq N_n + d, \quad R_n \leq \left\lfloor n^{\frac{1}{2(d+N_n)+1}} \right\rfloor, \quad R_n \leq \frac{\log(n)}{2\log(\log(N_n))} - \frac{1}{2}.$$

On $\mathcal{R}_n$, we have the following uniform upper bound for the convergence rate in deviation:

$$\lim_{C\to\infty} \limsup_{n\to\infty} \sup_{r \in (\mathcal{R}_n)^d} \sup_{f \in \mathcal{K}_r(L)} \mathbb{P}\left(D\left(f \| f_{\hat{\lambda}_n^*}\right) \geq \left(n^{-\frac{2\min(r)}{2\min(r)+1}} \right) C \right) = 0.$$

Therefore we obtained an estimator that is adaptive to the smoothness of the underlying density for an increasing set of regularity parameters without loss in the convergence rate. A simulation study confirms that the log-additive exponential model outperforms a truncated kernel estimator for the truncation model with various choices for the pdfs p_i, $i = 1, 2$. Figure 1.4 shows the true joint pdf and its estimators in the case when p_1 is the pdf of Normal mixture distribution and p_2 is the pdf of a Normal distribution.



(a) True density (b) LAESE (c) Kernel

Figure 1.4 – Joint density functions of the true density and its estimators with Normal mix marginals.

1.2.6 Perspectives

In the second part of the thesis, we considered the problem of nonparametric estimation of maximum entropy distributions of vectors of order statistics with support $\triangle$. When considering a distribution f with a support different from $\triangle$, problems may arise. Let f_i denote the i-th marginal pdf of of f and $A_i \subset \mathbb{R}$ its support for $1 \leq i \leq d$. When $A_i = A \subseteq \mathbb{R}$ for all $1 \leq i \leq d$, we can apply a strictly monotone mapping of A onto I to obtain a distribution with a product form supported on $\triangle$. The transformation needs to be chosen carefully in order to ensure that the resulting ℓ_i, $1 \leq i \leq d$ functions belong to certain Sobolev spaces. This can be particularly difficult when $A = \mathbb{R}$, where tail properties of the distribution have to be taken into consideration. When the supports A_i differ, there is no simple transformation that gives a random vector with joint pdf of the form (1.23). A possible way to treat this case consists of constructing a family of basis functions which has similar properties with respect to the support of f as the family $(\varphi_{i,k}, 1 \leq i \leq d, k \in \mathbb{N})$ with respect to $\triangle$. This would allow us to define an exponential model with this family of basis functions and support restricted to the support of f. A complete description for all types of supports could be subject for future work.

When we applied this estimation method to a real dataset in Chapter 7, we did not obtain satisfactory results, possibly due to the fact that the underlying joint pdf may not have the form (1.23). This calls for a statistical testing procedure which could determine, based on the available sample, whether the underlying joint pdf corresponds to the form (1.23) or not.

1.3 Industrial applications

The main motivation of this thesis work was to contribute to the need of modelling uncertainties in engineering studies related to probabilistic safety assessment. We first give a brief overview of the uncertainty quantification methodology developed by Électricité de France (EDF) Research and Development and related industrial research institutes. Then, we introduce the two case studies considered, which were presented at conferences dedicated to applied research in industrial context. The first case study, described in Section 1.3.2, concerns the simulation of physical parameters with monotonicity constraints in a numerical welding model. This study was presented at the 19^{th} Lambda-Mu Conference in Dijon, see [31]. The second case study, detailed in Section 1.3.3, proposes a nonparametric method to estimate the joint density of the dimensions of flaws in a mechanical component in a power plant. This work was presented at the 25^{th} European Safety and Reliability Conference in Zürich, see [34].

1.3.1 Uncertainty quantification in engineering : the EDF framework

In this section we present the general uncertainty treatment methodology of EDF Research and Development, as well as a dedicated software platform called OpenTURNS which implements the methods related to uncertainty treatment.

A four-step iterative method

Recently the need to comply with regulatory requirements led EDF to develop a global methodology framework for treating uncertainties for models and simulations in collaboration with other major companies, industrial research institutions and academic partners. We refer to [138] and [137] for an overview on this topic with recent developments and numerous examples. The resulting generic uncertainty treatment methodology consists of a four step process illustrated by Figure 1.5. These steps are the following:

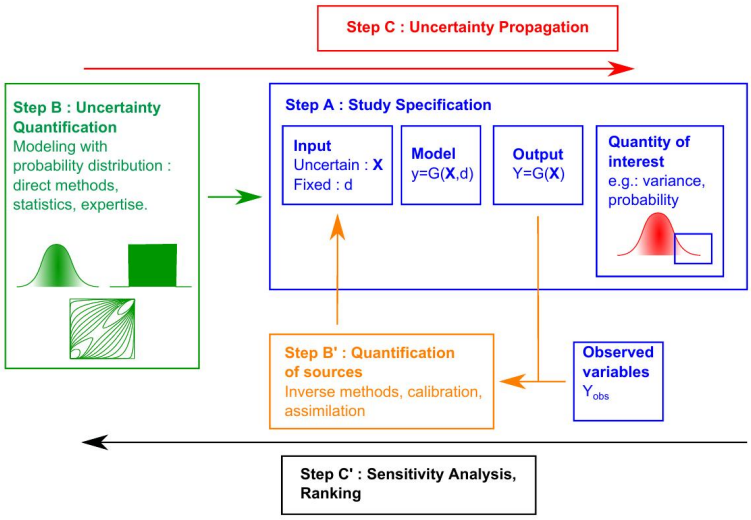


Figure 1.5 – The uncertainty treatment methodology. Source: [15].

Step A: Uncertainty Problem Specification. The first step consists of specifying the model we would like to study as well as the sources of uncertainty. The model (which can be an analytical formula as well as a computer code or an experimental process) is represented as a function $G : \mathbb{R}^{p+q} \mapsto \mathbb{R}^z$. We can split the input vector into a set of parameters $X = (X_1, \ldots, X_p)$ that are subject to uncertainties, and a set of parameters $d = (d_1, \ldots, d_q)$ that are considered as fixed. The output Y becomes a z-dimensional random vector:

$$Y = G(X, d),$$

where $X \in \mathbb{R}^p$ is a random vector representing the uncertain inputs and $d \in \mathbb{R}^q$ is the vector of deterministic inputs. We also need to specify the so called *quantity of interest*, a relevant feature of the output variable Y which we would like to study in order to answer the regulatory demand. Such features can be measures of central tendency (mean, median or mode), measures of dispersion (range, variance or standard deviation), the probability of exceeding certain thresholds, etc.

Step B: Uncertainty Quantification of the Input. Once the sources of uncertainty have been fixed, we need to propose a probabilistic model to account for them. This involves determining the joint distribution of the random vector X. The modelling procedure depends on the available information. This information can come in the form of an expert judgement, an available sample or some physical constraints which need to be respected. If there is only scarce information based on expert opinions, a common method is to use a the Maximum Entropy Principal (originating from [161]) to propose a distribution which is the least informative given the available expert knowledge. If we possess a data set of sufficient size, we can identify a parametric model and estimate its parameters or we can proceed with a non-parametric approach if a convenient parametric model does not emerge. In all cases, the probabilistic model needs to be validated before we continue our analysis.

Step C: Uncertainty Propagation. With the probabilistic model for the random input vector established, uncertainties are propagated to the model outcome and estimate the quantity of interest. Depending on the complexity and cost of the model evaluation, several methods can be applied. The easiest case is when analytical formulas are available. When this is not the case, we have to rely on approximation techniques. If we are interested in central tendency or dispersion measures, we can use Monte Carlo sampling methods. If the evaluation of the model G is costly, then we can apply the Taylor variance decomposition method, or build a metamodel, whose cost is negligible compared to the cost of G, and perform a Monte Carlo study with it. For the probability of threshold exceedance, especially when we are interested in rare events, several techniques exist to accelerate the estimation such as importance sampling, subset sampling, FORM-SORM methods, etc.

Step C': Analysis of Sensitivity and Importance Ranking. Finally, we analyse the sensitivity of the quantity of interest with respect to the input variables in order to determine which variables have the most influence on this quantity. Sensitivity can be measured by simple correlation coefficients or variance based methods such as Sobol's indices. The ranking can help us to determine which variables require further attention in order to obtain more precise results after taking some feedback action and refining our model. Then the procedure can be iterated to attain more satisfactory results.

The work presented in this thesis concerns Step B of this scheme, the probabilistic modelling of random input variables. The currently accepted EDF approach consists of modelling the joint distribution of the input random vector $X = (X_1, \ldots, X_p)$ by establishing the marginal distributions X_i, $1 \leq i \leq p$, and explicitly modelling the dependence structure via giving the connecting copula of X. Chapter 3 considers the case when the expert's knowledge include the precise characterization of the marginal distributions (cumulative distribution function fixed) and an almost sure ordering constraint between the components of X. The Maximum Entropy Principal is applied to obtain the distribution which contains the least information in addition to these constraints. In the first application case of Section 1.3.2, where physical parameters which are monotone functions of the temperature are modelled, the setup was similar to this. In absence of significant amount of empirical data, the marginals were considered to be uniform random variables on intervals which correspond to the scarce experimental results, and the physics of the welding model imposes the ordering constraint.

The second application case of Section 1.3.3 studies a situation where a sample is available from the random input vector X, exhibiting similar ordering constraints as required in Chapter 3. We attempt to model the random vector X with the family of maximum entropy distributions of ordered random vector obtained in this section, and propose a non-parametric estimation method in Chapter 6 to estimate the joint distribution of X. We compare the properties of this new model to models proposed in previous papers (for example [148]) in an uncertainty quantification study concerning the evaluation of the probability of a threshold exceedance.

OpenTURNS : an industrial software for uncertainty treatment

To implement the uncertainty treatment methodology described in the previous section, EDF and its industrial partners Airbus Group, Phimeca Engineering and IMACS developed a dedicated open source software platform named OpenTURNS. This C++ library endowed with a Python TUI, available to download at `www.openturns.org`, implements the step-by-step framework with a large selection of available methods for each step of the uncertainty treatment methodology based on the specific characteristics of the underlying problem. In particular, the methods cited in the description of methodology are readily implemented, see [15] for a comprehensive (but not exhaustive) overview of the functionalities.

One of the main innovative features of the OpenTURNS platform is the ability of modelling multivariate distributions by copula functions. This allows us to define the joint multivariate cumulative distribution function of the random vector X of uncertain inputs as the composition

of a copula function with the univariate marginal cumulative distribution functions. With a large selection of copulas, including most well known parametric families as well as the ability to combine different copulas via composition or to extract the copula of an arbitrary multivariate distribution, OpenTURNS gives the possibility to specify complex dependence structures as opposed to a simplistic, and often severely wrong, assumption of independence of the input variables.

In particular, the maximum entropy distribution of order statistics associated to a vector of marginals $(F_i, 1 \leq i \leq p)$ given in Chapter 3 was implemented in this platform. One can define such a distribution using the *MaximumEntropyOrderStatisticsDistribution()* object by providing the list of marginal cumulative distribution functions. Then OpenTURNS verifies the compatibility conditions of stochastic ordering between them before building the joint cumulative distribution function. The following code implements the case when $p = 2$, and the marginals are normally distributed with unit variance and means 0 and 1, respectively.

```
marg_1 = Normal(0,1)
marg_2 = Normal(1,1)
list_m = [marg_1,marg_2]
max_entr_dist = MaximumEntropyOrderStatisticsDistribution(list_m)
```

It is also possible to only define the copula function of the maximum entropy distribution of order statistics through the object *MaximumEntropyOrderStatisticsCopula()*, and then combine it with any set of marginals to obtain a joint distribution. Figure 1.3 illustrate the joint density and the copula of the obtained distribution.

The non-parametric estimations of Chapters 6 and 7 have also been carried out with the help of OpenTURNS. Many aspects of the library were utilized in the code such as the readily available families of orthogonal polynomials, sophisticated multidimensional numerical integration, the TNC (Truncated Newton Constrained) optimization algorithm or the multivariate kernel density estimator. An eventual implementation of the estimation method of Chapter 6 could enrich the library with further non-parametric methods besides the kernel density estimator.

1.3.2 Numerical model for welding

The first application case study considered during my thesis work considers the probabilistic modelling of input parameters of a numerical welding simulation procedure in order to conduct uncertainty quantification analysis. The simulation is based on a numerical finite element method based on a thermal-mechanical model. The simulations are used to evaluate the characteristics of residual stresses formed during the welding procedure, which can severely impact the lifetime of the welded component. Input parameters include several physical characteristics of the material such as Young's modulus, yield strength, coefficient of thermal expansion, etc. These parameters are evaluated at different temperatures covering a large range. While the values of the parameters for low temperatures are relatively well known, data is very scarce on their values for high temperatures, which leads to uncertainties.

Several sensitivity analysis studies have been carried out by researchers at EDF Research and Development in order to assess the influence of the input parameters on the appearance of residual stresses, see [7], [144] and [145]. These papers consider the welding of steel plates (focusing on the type 316L stainless steel), and the thermal-mechanical calculations are implemented using the software platform *Code_Aster* developed by EDF Research and Development.

The contributions of this thesis work is presented in Chapter 4. We propose an alternative method to the currently utilized approach for the modelling of physical parameter profiles which are monotone functions of the temperature. The current approach enforces the monotonicity constraint by imposing a function for the mean values at each temperature, and introducing variability by adding an error function multiplied by zero-mean noise. There are multiple drawbacks of this approach. First, it implies a strong hypothesis on the form of the function curve, which is not supported by the available empirical data. Second, it can also result in parameter profiles which are not monotone as a function of the temperature. Third, since the uncertainty

is modelled by a single random noise, the marginal distributions of the parameter values will be the same up to an affine transformation.

Several papers signed by researchers of the Industrial Risk Management Department addressed the question of generating samples of parameters with ordering and marginal constraints. [140] proposes a constrained Latin Hypercube Sampling method to generate pseudo-random realizations of the parameters. The drawbacks of this method is that the exact joint distribution of the generated sample is unknown, therefore no control on the convergence of the Monte Carlo simulation is available. In [119], the authors give copula-based construction of a multivariate joint distribution which verifies both constraints. However the analytic formula of this copula requires the computation of a set of functions which are solutions of ordinary differential equations, which are generally hard to find.

The method we suggest is to use the maximum entropy distribution of order statistics proposed by Chapter 3, see formula (1.10), to define the joint distribution of the parameter values at different temperatures. This method has multiple advantages over the previously discussed approaches. A simple analytic formula is available for the joint density function, which almost surely verifies the ordering constraint as opposed to the currently maintained approach. The maximum entropy distribution also has a clear information-theory interpretation. Furthermore, the marginals can be freely chosen for each temperature value (as long as they are stochastically ordered), and there is no function form enforced. Chapter 4 also discusses in detail the case of uniform marginals, with an inversion method with explicit formulas provided for elementary simulation.

1.3.3 Modelling physical flaws in a passive mechanic component

The second application case concerns the probabilistic modelling of input parameters for a fissure propagation simulation code. This code implements the physical model of propagation of initial cracks in a metallic material under severe pressure. If the calculated intensity factor exceeds de resistance of the steel (which also depends on the input parameters), a brutal rupture may occur, potentially damaging the integrity of the examined component. We focus on the joint modelling of two inputs, the initial depth and the length of the crack. These variables naturally exhibit dependence.

Contrary to the previous case of Section 1.3.2, a data set of several hundred joint observations is available to aid the modelling process. The data comes from either observations accumulated during regular inspection of components operating in EDF power plants, or from controlled experiments. The data suggest that the dimensions verify the ordering constraint, as the length of the cracks is greater then the depth of the crack for all observations (a physical argument to support this hypothesis is yet to emerge).

The currently accepted approach makes the assumption that the depth and the ratio between the depth and length are independent, and a distribution is fitted with parametric families suitably chosen for both variables. In an attempt to correctly model the dependence structure between the dimensions, [148] considers the estimation of the copula for the two-dimensional distribution by parametric families such as Gaussian, Frank or Gumbel copulas. The inconvenience of this approach is that the parametric families contain only symmetric copulas, therefore potential asymmetric features of the dependence structure can not be accounted for. In particular, if the hypothesis on the almost sure ordering relationship between the dimensions is correct, then its copula has a restricted support and is not symmetric, as it was seen in Chapter 3.

In the conference paper [34], we presented the possibility of using the non-parametric estimation procedure of Chapter 6 to estimate the joint density of the dimensions of the cracks. We keep in mind that even if the model does not fit as well to the empirical data as other models, preference should be given to a model which is conservative, i.e. which potentially overestimates the risk of a brutal rupture.

The results of the study show that even though the proposed non-parametric estimator disperses the probabilistic mass on a larger domain than the previous models, it considerably underestimates the failure probability compared to the other models. This is due to the fact that

a failure is more likely to occur when both dimensions assume large values, whereas the non-parametric estimator gives considerable weight to combinations with a large value for the length, but smaller value for the depth. This may indicate that we overlooked some physical phenomenon by considering only ordered random vectors. An improvement to the modelling procedure would be to consider maximum entropy copulas of order statistics relative to a reference copula with high upper tail-dependence coefficient instead of the independent copula, so that more penalizing scenarios are accounted for with higher probability.

Part I

Modelling the dependence structure of order statistics: a copula theory approach

Chapter 2

Maximum entropy copula with given diagonal section

2.1 Introduction

Dependence of random variables can be described by copula distributions. A copula is the cumulative distribution function of a random vector $U = (U_1, \ldots, U_d)$ with U_i uniformly distributed on $I = [0, 1]$. For an exhaustive overview on copulas, we refer to Nelsen [132]. The diagonal section δ of a d-dimesional copula C, defined on I as $\delta(t) = C(t, \ldots, t)$ is the cumulative distribution function of $\max_{1 \leq i \leq d} U_i$. The function δ is non-decreasing, d-Lipschitz, and verifies $\delta(t) \leq t$ for all $t \in I$ with $\delta(0) = 0$ and $\delta(1) = 1$. It was shown that if a function δ satisfies these properties, then there exists a copula with δ as diagonal section (see Bertino [20] or Fredricks and Nelsen [78] for $d = 2$ and Cuculescu and Theodorescu [49] for $d \geq 2$).

Copulas with a given diagonal section have been studied in different papers, as the diagonal sections are considered in various fields of application. Beyond the fact that δ is the cumulative distribution function of the maximum of the marginals, it also characterizes the tail dependence of the copula (see Joe [105] p.33. and references in Nelsen et al. [134], Durante and Jaworski [68], Jaworski [102]) as well as the generator for Archimedean copulas (Sungur and Yang [164]). For $d = 2$, Bertino in [20] introduces the so-called Bertino copula B_δ given by $B_\delta(u, v) = u \wedge v - \min_{u \wedge v \leq t \leq u \vee v}(t - \delta(t))$ for $u, v \in I$. Fredricks and Nelsen in [78] give the example called diagonal copula defined by $K_\delta(u, v) = \min(u, v, (\delta(u) + \delta(v))/2)$ for $u, v \in I$. In Nelsen et al. [133, 134] lower and upper bounds related to the pointwise partial ordering are given for copulas with a given diagonal section. They showed that if C is a symmetric copula with diagonal section δ, then for every $u, v \in I$, we have:

$$B_\delta(u, v) \leq C(u, v) \leq K_\delta(u, v).$$

Durante et al. [69] provide another construction of copulas for a certain class of diagonal sections, called MT-copulas named after Mayor and Torrens and defined as $D_\delta(u, v) = \max(0, \delta(x \vee y) - |x - y|)$. Bivariate copulas with given sub-diagonal sections $\delta_{x_0} : [0, 1 - x_0] \to [0, 1 - x_0], \delta_{x_0}(t) = C(x_0 + t, t)$ are constructed from copulas with given diagonal sections in Quesada-Molina et al. [147]. Durante et al. [70] and [134] introduce the technique of diagonal splicing to create new copulas with a given diagonal section based on other such copulas. According to [68] for $d = 2$ and Jaworski [102] for $d \geq 2$, there exists an absolutely continuous copula with diagonal section δ if and only if the set $\Sigma_\delta = \{t \in I; \delta(t) = t\}$ has zero Lebesgue measure. de Amo et al. [57] is an extension of [68] for given sub-diagonal sections. Further construction of possibly asymmetric absolutely continuous bidimensional copulas with a given diagonal section is provided in Erdely and González [76].

Our aim is to find the most uninformative copula with a given diagonal section δ. We choose here to maximize the relative entropy to the uniform distribution on I^d, among the copulas with given diagonal section. This is equivalent to minimizing the Kullback-Leibler divergence with respect to the independent copula. The Kullback-Leibler divergence is finite only for absolutely

continuous copulas. The previously introduced bivariate copulas B_δ, K_δ and D_δ are not absolutely continuous, therefore their Kullback-Leibler divergence is infinite. Possible other entropy criteria, such as Rényi, Tsallis, etc. are considered for example in POUGAZA AND MOHAMMAD-DJAFARI [146]. We recall that the entropy of a d-dimensional absolutely continuous random vector $X = (X_1, \dots, X_d)$ can be decomposed as the sum of the entropy of the marginals and the entropy of the corresponding copula (see ZHAO AND LIN [183]) :

$$H(X) = \sum_{i=1}^{d} H(X_i) + H(U),$$

where $H(Z) = -\int f_Z(z) \log f_Z(z) dz$ is the entropy of the random variable Z with density f_Z, and $U = (U_1, \dots, U_d)$ is a random vector with U_i uniformly distributed on I, such that U has the same copula as X; namely U is distributed as $\left(F_1^{-1}(X_1), \dots F_d^{-1}(X_d)\right)$ with F_i the cumulative distribution function of X_i. Maximizing the entropy of X with given marginals therefore corresponds to maximizing the entropy of its copula. The maximum relative entropy approach for copulas has an extensive litterature. Existence results for an optimal solution on convex closed subsets of copulas for the total variation distance can be derived from CSISZÁR [48]. A general discussion on abstract entropy maximization is given by BORWEIN ET AL. [25]. This theory was applied for copulas and a finite number of expectation constraints in BEDFORD AND WILSON [16]. Some applications for various moment-based constraints include rank correlation (MEEUWISSEN AND BEDFORD [128], CHU [46], PIANTADOSI ET AL. [141]) and marginal moments (PASHA AND MANSOURY [139]).

We shall apply the theory developed in [25] to compute the density of the maximum entropy copula with a given diagonal section. We show that there exists a copula with diagonal section δ and finite entropy if and only if δ satisfies: $\int_I |\log(t - \delta(t))| dt < +\infty$. Notice that this condition is stronger than the condition of Σ_δ having zero Lebesgue measure which is required for the existence of an absolutely continuous copula with diagonal section δ. Under this condition, and in the case of $\Sigma_\delta = \{0, 1\}$, the optimal copula's density c_δ turns out to be of the form, for $x = (x_1, \dots, x_d) \in I^d$:

$$c_\delta(x) = b(\max(x)) \prod_{x_i \neq \max(x)} a(x_i),$$

with the notation $\max(x) = \max_{1 \leq i \leq d} x_i$, see Proposition 2.4. The optimal copula's density in the general case is given in Theorem 2.5. Notice that c_δ is symmetric: it is invariant under the permutation of the variables. This provides a new family of absolutely continuous symmetric copulas with given diagonal section enriching previous work on this subject that we discussed, see [20],[68],[70],[69],[76],[78],[134]. We also calculate the maximum entropy copula for diagonal sections that arise from well-known families of bivariate copulas.

The rest of the paper is organised as follows. Section 2.2 introduces the definitions and notations used later on, and gives the main theorems of the paper. In Section 2.3 we study the properties of the feasible solution c_δ of the problem for a special class of diagonal sections with $\Sigma_\delta = \{0, 1\}$. In Section 2.4, we formulate our problem as an optimization problem with linear constraints in order to apply the theory established in [25]. Then in Section 2.5 we give the proof for our main theorem showing that c_δ is indeed the optimal solution when $\Sigma_\delta = \{0, 1\}$. In Section 2.6 we extend our results for the general case when Σ_δ has zero Lebesgue measure. We give in Section 2.7 several examples with diagonals of popular bivariate copula families such as the Gaussian, Gumbel or Farlie-Gumbel-Morgenstern copulas among others. In the Gaussian case, we illustrate how different the Gaussian copula and the corresponding maximum entropy copula can be, by calculating conditional extreme event probabilities.

2.2 Main results

Let $d \geq 2$ be fixed. We recall a function C defined on I^d, with $I = [0, 1]$, is a d-dimensional copula if there exists a random vector $U = (U_1, \dots, U_d)$ such that U_i are uniform on I and

$C(u) = \mathbb{P}(U \leq u)$ for $u \in I^d$, with the convention that $x \leq y$ for $x = (x_1, \ldots x_d)$ and $y = (y_1, \ldots, y_d)$ elements of $\mathbb{R}^d$ if and only if $x_i \leq y_i$ for all $1 \leq i \leq d$. We shall say that C is the copula of U. We refer to [132] for a monograph on copulas. The copula C is said absolutely continuous if the random variable U has a density, which we shall denote by c_C. In this case, we have for all $u \in I^d$:

$$C(u) = \int_{I^d} c_C(v) \mathbf{1}_{\{v \leq u\}} \, dv.$$

When there is no confusion, we shall write c for the density c_C associated to the copula C. We denote by $\mathcal{C}$ the set of d-dimensional copulas and by $\mathcal{C}_0$ the subset of the d-dimensional absolutely continuous copulas.

The diagonal section δ_C of a copula C is defined by: $\delta_C(t) = C(t, \ldots, t)$. Let us note, for $u \in \mathbb{R}^d$, $\max(u) = \max_{1 \leq i \leq d} u_i$. Notice that if C is the copula of U, then δ_C is the cumulative distribution function of $\max(U)$ as $\delta_C(t) = \mathbb{P}(\max(U) \leq t)$ for $t \in I$. We denote by $\mathcal{D} = \{\delta_C, C \in \mathcal{C}\}$ the set of diagonal sections of d-dimensional copulas and by $\mathcal{D}_0 = \{\delta_C; C \in \mathcal{C}_0\}$ the set of diagonal sections of absolutely continuous copulas. According to [78], a function δ defined on I belongs to $\mathcal{D}$ if and only if:

(i) δ is a cumulative function on $[0, 1]$: $\delta(0) = 0$, $\delta(1) = 1$ and δ is non-decreasing;

(ii) $\delta(t) \leq t$ for $t \in I$ and δ is d-Lipschitz: $|\delta(s) - \delta(t)| \leq d\,|s - t|$ for $s, t \in I$.

For $\delta \in \mathcal{D}$, we shall consider the set $\mathcal{C}^\delta = \{C \in \mathcal{C}; \delta_C = \delta\}$ of copulas with diagonal section δ, and the subset $\mathcal{C}_0^\delta = \mathcal{C}^\delta \bigcap \mathcal{C}_0$ of absolutely continuous copulas with section δ. According to [68] and [102], the set $\mathcal{C}_0^\delta$ is non empty if and only if the set $\Sigma_\delta = \{t \in I; \delta(t) = t\}$ has zero Lebesgue measure.

For a non-negative measurable function f defined on $I^k, k \in \mathbb{N}$, we set

$$\mathcal{I}_k(f) = \int_{I^k} f(x) \log(f(x)) \, dx,$$

with the convention $0 \log(0) = 0$. Since copulas are cumulative functions of probability measures, we will consider the Kullback-Leibler divergence relative to the uniform distribution as a measure of entropy, see [48]:

$$\mathcal{I}(C) = \begin{cases} \mathcal{I}_d(c) & \text{if } C \in \mathcal{C}_0, \\ +\infty & \text{if } C \notin \mathcal{C}_0, \end{cases}$$

with c the density associated to C when $C \in \mathcal{C}_0$. Notice that the Shannon-entropy introduced in [161] of the probability measure P defined on I^d with cumulative distribution function C is defined as $H(P) = -\mathcal{I}(C)$. Thus minimizing the Kullback-Leibler divergence $\mathcal{I}$ (w.r.t. the uniform distribution) is equivalent to maximizing the Shannon-entropy. It is well known that the copula Π with density $c_\Pi = 1$, which corresponds to $(U_i, 0 \leq i \leq d)$ being independent, minimizes $\mathcal{I}(C)$ over $\mathcal{C}$.

We shall minimize the Kullback-Leibler divergence $\mathcal{I}$ over the set $\mathcal{C}^\delta$ or equivalently over $\mathcal{C}_0^\delta$ of copulas with a given diagonal section $\delta \in \mathcal{D}$ (in fact for $\delta \in \mathcal{D}_0$ as otherwise $\mathcal{C}_0^\delta$ is empty). If C minimizes $\mathcal{I}$ on $\mathcal{C}^\delta$, it means that C is the least informative (or the "most random") copula with given diagonal section δ.

For $\delta \in \mathcal{D}$, let us denote:

$$\mathcal{J}(\delta) = \int_I |\log(t - \delta(t))| \, dt. \tag{2.1}$$

Notice that $\mathcal{J}(\delta) \in [0, +\infty]$ and it is infinite if $\delta \notin \mathcal{D}_0$. Since δ is d-Lipschitz, the derivative δ' of δ exists a.e. and since δ is non-decreasing we have a.e. $0 \leq \delta' \leq d$. This implies that $\mathcal{I}_1(\delta')$ and $\mathcal{I}_1(d - \delta')$ are well defined. Let us denote:

$$\mathcal{G}(\delta) = \mathcal{I}_1(\delta') + \mathcal{I}_1(d - \delta') - d \log(d) - (d - 1). \tag{2.2}$$

Since for any function f such that $0 \leq f \leq d$ we have $-1/\mathrm{e} \leq \mathcal{I}_1(f) \leq d \log(d)$, we can give a rough upper bound for $|\mathcal{G}(\delta)|$:

$$\sup_{\delta \in \mathcal{D}} |\mathcal{G}(\delta)| \leq d + d \log(d). \tag{2.3}$$

For $\delta \in \mathcal{D}_0$ with $\Sigma_\delta = \{0,1\}$, we define the function c_δ as:

$$c_\delta(x) = b(\max(x)) \prod_{x_i \neq \max(x)} a(x_i) \qquad \text{for a.e. } x = (x_1, \dots, x_d) \in I^d, \tag{2.4}$$

where the functions a and b are given by, for $r \in I$:

$$a(r) = \frac{d - \delta'(r)}{d} h(r)^{-1+1/d} \, \mathrm{e}^{F(r)} \quad \text{and} \quad b(r) = \frac{\delta'(r)}{d} h(r)^{-1+1/d} \, \mathrm{e}^{-(d-1)F(r)}, \tag{2.5}$$

with h and F defined as:

$$h(r) = r - \delta(r), \quad F(r) = \frac{d-1}{d} \int_{\frac{1}{2}}^{r} \frac{1}{h(s)} \, ds. \tag{2.6}$$

Remark 2.1. Notice that we define F in (2.6) as an integral from $1/2$ to r. However, the value $1/2$ can be chosen arbitrarily on $(0,1)$ as it will not affect the definition of the function c_δ in (2.4).

The following Proposition shows that c_δ is an absolutely continuous copula whose diagonal section is δ. The proof of this Proposition can be found in Section 2.3 and Section 2.8.1 is dedicated to the proof of (2.7).

Proposition 2.2. *Let $\delta \in \mathcal{D}_0$ with $\Sigma_\delta = \{0,1\}$. The function c_δ given by (2.4) is the density of a symmetric copula C_δ with diagonal section δ.*

Furthermore, we have:

$$\mathcal{I}(C_\delta) = (d-1)\mathcal{J}(\delta) + \mathcal{G}(\delta). \tag{2.7}$$

This and (2.3) readily implies the following Remark.

Remark 2.3. Let $\delta \in \mathcal{D}_0$ such that $\Sigma_\delta = \{0,1\}$. We have $\mathcal{I}(C_\delta) < +\infty$ if and only if $\mathcal{J}(\delta) < +\infty$.

We can now state our main result in the simpler case $\Sigma_\delta = \{0,1\}$. It gives the necessary and sufficient condition for C_δ to be the unique optimal solution of the minimization problem. The proof is given in Section 2.5.

Proposition 2.4. *Let $\delta \in \mathcal{D}_0$ such that $\Sigma_\delta = \{0,1\}$.*

a) If $\mathcal{J}(\delta) = +\infty$ then $\min_{C \in \mathcal{C}^\delta} \mathcal{I}(C) = +\infty$.

b) If $\mathcal{J}(\delta) < +\infty$ then $\min_{C \in \mathcal{C}^\delta} \mathcal{I}(C) < +\infty$ and C_δ is the unique copula such that $\mathcal{I}(C_\delta) = \min_{C \in \mathcal{C}^\delta} \mathcal{I}(C)$.

To give the answer in the general case where Σ_δ has zero Lebesgue measure, which is the necessary and sufficient condition for $\mathcal{C}_0^\delta \neq \emptyset$, we need some extra notations. Since δ is continuous, we get that $I \setminus \Sigma_\delta$ can be written as the union of non-empty open disjoint intervals $((\alpha_j, \beta_j), j \in J)$, with $\alpha_j < \beta_j$ and J at most countable. Notice that $\delta(\alpha_j) = \alpha_j$ and $\delta(\beta_j) = \beta_j$. For $J \neq \emptyset$ and $j \in J$, we set $\Delta_j = \beta_j - \alpha_j$ and for $t \in I$:

$$\delta^j(t) = \frac{\delta(\alpha_j + t\Delta_j) - \alpha_j}{\Delta_j}. \tag{2.8}$$

It is clear that δ^j satisfies (i) and (ii) and it belongs to $\mathcal{D}_0$ as $\Sigma_{\delta^j} = \{0,1\}$. Let c_{δ^j} be defined by (2.4) with δ replaced by δ^j. For $\delta \in \mathcal{D}_0$ such that $\Sigma_\delta \neq \{0,1\}$, we define the function c_δ by, for $u \in I^d$:

$$c_\delta(u) = \sum_{j \in J} \frac{1}{\Delta_j} c_{\delta^j} \left(\frac{u - \alpha_j \mathbf{1}}{\Delta_j} \right) \mathbf{1}_{(\alpha_j, \beta_j)^d}(u), \tag{2.9}$$

with $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^d$. It is easy to check that c_δ is a copula density and that is zero outside $[\alpha_j, \beta_j]^d$ for $j \in J$. We state our main result in the general case whose proof is given in Section 2.6.

Theorem 2.5. *Let $\delta \in \mathcal{D}$.*

a) If $\mathcal{J}(\delta) = +\infty$ then $\min_{C\in\mathcal{C}^\delta} \mathcal{I}(C) = +\infty$.

b) If $\mathcal{J}(\delta) < +\infty$ then $\min_{C\in\mathcal{C}^\delta} \mathcal{I}(C) < +\infty$ and there exists a unique copula $C_\delta \in \mathcal{C}^\delta$ such that $\mathcal{I}(C_\delta) = \min_{C\in\mathcal{C}^\delta} \mathcal{I}(C)$. Furthermore, we have:

$$\mathcal{I}(C_\delta) = (d-1)\mathcal{J}(\delta) + \mathcal{G}(\delta);$$

the copula C_δ is absolutely continuous, symmetric; its density c_δ is given by (2.4) if $\Sigma_\delta = \{0,1\}$ or by (2.9) if $\Sigma_\delta \neq \{0,1\}$.

Remark 2.6. For $\delta \in \mathcal{D}$, notice the condition $\mathcal{J}(\delta) < +\infty$ implies that Σ_δ has zero Lebesgue measure, and therefore, according to [68] and [102], $\delta \in \mathcal{D}_0$. And if $\delta \notin \mathcal{D}_0$, then $\mathcal{I}(C) = +\infty$ for all $C \in \mathcal{C}^\delta$. Therefore, we could replace the condition $\delta \in \mathcal{D}$ by $\delta \in \mathcal{D}_0$ in Theorem 2.5.

2.3 Proof of Proposition 2.2

We assume that $\delta \in \mathcal{D}_0$ and $\Sigma_\delta = \{0,1\}$. We give the proof of Proposition 2.2, which states that C_δ, with density c_δ given by (2.4), is indeed a symmetric copula with diagonal section δ whose entropy is given by (2.7).

Recall the definition of h, F, a, b and c_δ given by (2.4) to (2.6). Notice that by construction c_δ is non-negative and well defined on I^d. In order to prove that c_δ is the density of a copula, we only have to prove that for all $1 \leq i \leq d$, $r \in I$:

$$\int_{I^d} c_\delta(u)\mathbf{1}_{\{u_i\leq r\}}\, du = r,$$

or equivalently

$$\int_{I^d} c_\delta(u)\mathbf{1}_{\{u_i\geq r\}}\, du = 1 - r.$$

We define for $r \in I$:

$$A(r) = \int_0^r a(t)\, dt. \tag{2.10}$$

Elementary computations yield for $r \in (0,1)$:

$$A(r) = h^{1/d}(r)\, \mathrm{e}^{F(r)}. \tag{2.11}$$

Notice that $F(0) \in [-\infty, 0]$ which implies that $A(0) = 0$. A direct integration gives:

$$d\int_I A^{d-1}(s)b(s)\mathbf{1}_{\{s\geq r\}}\, ds = 1 - \delta(r). \tag{2.12}$$

We also have:

$$\begin{aligned}(d-1)\int_I A^{d-2}(s)b(s)\mathbf{1}_{\{s\geq r\}}\, ds &= \frac{(d-1)}{d}\int_I \delta'(s)h^{-1/d}(s)\,\mathrm{e}^{-F(s)}\,\mathbf{1}_{\{s\geq r\}}\, ds \\ &= \left[-h^{1-1/d}(s)\,\mathrm{e}^{-F(s)}\right]_{s=r}^{1} \\ &= h^{1-1/d}(r)\,\mathrm{e}^{-F(r)},\end{aligned} \tag{2.13}$$

where we used for the last step that $h(1) = 0$ and $F(1) \in [0, \infty]$. We have:

$$\begin{aligned}
\int_{I^d} c_\delta(u)\mathbf{1}_{\{u_i \geq r\}}\, du &= \int_{I^d} b(\max(u)) \prod_{u_j \neq \max(u)} a(u_j)\mathbf{1}_{\{u_i \geq r\}}\, du \\
&= \int_I A^{d-1}(s)b(s)\mathbf{1}_{\{s \geq r\}}\, ds \\
&\qquad + (d-1)\int_I A^{d-2}(s)b(s)(A(s) - A(r))\mathbf{1}_{\{s \geq r\}}\, ds \\
&= d\int_I A^{d-1}(s)b(s)\mathbf{1}_{\{s \geq r\}}\, ds \\
&\qquad - (d-1)A(r)\int_I A^{d-2}(s)b(s)\mathbf{1}_{\{s \geq r\}}\, ds \\
&= 1 - \delta(r) - h(r) \\
&= 1 - r,
\end{aligned}$$

where in the second equality we separated the integral according to $\max(u) = u_i$ or not and used (2.10), then in the fourth equality we used (2.12) and (2.13). This implies that c_δ is indeed the density of a copula. We denote by C_δ the copula with density c_δ. We check that δ is the diagonal section of C_δ. Using (2.12), we get, for $r \in I$:

$$\begin{aligned}
\int_{I^d} c_\delta(u)\mathbf{1}_{\{\max(u) \leq r\}}\, du &= \int_{I^d} b(\max(u)) \prod_{u_i \neq \max(u)} a(u_i)\mathbf{1}_{\{\max(u) \leq r\}}\, du \\
&= d\int_I A^{d-1}(s)b(s)\mathbf{1}_{\{s \leq r\}}\, ds \\
&= \delta(r).
\end{aligned}$$

The calculations which show that the entropy of C_δ is given by (2.7) can be found in Section 2.8.1.

2.4 The minimization problem

Let $\delta \in \mathcal{D}_0$. As a first step we will show, using [25], that the problem of a maximum entropy copula with a given diagonal section δ has at most a unique optimal solution. To formulate this problem in the framework of [25], we introduce the continuous linear functional $\mathcal{A} = (\mathcal{A}_i, 1 \leq i \leq d+1) : L^1(I^d) \to L^1(I)^{d+1}$ defined by, for $1 \leq i \leq d$, $f \in L^1(I^d)$ and $r \in I$,

$$\mathcal{A}_i(f)(r) = \int_{I^d} f(u)\mathbf{1}_{\{u_i \leq r\}}\, du, \quad \text{and} \quad \mathcal{A}_{d+1}(f)(r) = \int_{I^d} f(u)\mathbf{1}_{\{\max(u) \leq r\}}\, du.$$

We also define $b^\delta = (b_i, 1 \leq i \leq d+1) \in L^1(I)^{d+1}$ with $b_{d+1} = \delta$ and $b_i = \mathrm{id}_I$ for $1 \leq i \leq d$, with id_I the identity map on I. Notice that the conditions $\mathcal{A}_i(c) = b_i$, $1 \leq i \leq d$, and $c \geq 0$ a.e. imply that c is the density of a copula $C \in \mathcal{C}_0$. If we assume further that the condition $\mathcal{A}_{d+1}(c) = b_{d+1}$ holds then the diagonal section of C is δ (thus $C \in \mathcal{C}_0^\delta$).

Since $\mathcal{I}$ is infinite outside $\mathcal{C}_0^\delta$ and the density of any copula in $\mathcal{C}_0$ belongs to $L^1(I^d)$, we get that minimizing $\mathcal{I}$ over $\mathcal{C}^\delta$ is equivalent to the optimization problem (P^δ) given by:

$$\text{minimize } \mathcal{I}_d(c) \text{ subject to } \begin{cases} \mathcal{A}(c) = b^\delta, \\ c \geq 0 \text{ a.e. and } c \in L^1(I^d). \end{cases} \qquad (P^\delta)$$

We say that a function f is feasible for (P^δ) if $f \in L^1(I^d)$, $f \geq 0$ a.e., $\mathcal{A}(f) = b^\delta$ and $\mathcal{I}_d(f) < +\infty$. Notice that any feasible f is the density of a copula. We say that f is an optimal solution to (P^δ) if f is feasible and $\mathcal{I}_d(f) \leq \mathcal{I}_d(g)$ for all g feasible.

Proposition 2.7. *Let $\delta \in \mathcal{D}$. If there exists a feasible c, then there exists a unique optimal solution to (P^δ) and it is symmetric.*

Proof. Since $\mathcal{A}(f) = b^\delta$ implies $\mathcal{A}_1(f)(1) = b_1(1)$ that is $\int_{I^d} f(x)\,dx = 1$, we can directly apply Corollary 2.3 of [25] which states that if there exists a feasible c, then there exists a unique optimal solution to (P^δ). Since the constraints are symmetric and the functional $\mathcal{I}_d$ is also symmetric, we deduce that the unique optimal solution is also symmetric. □

The next Proposition gives that the set of zeros of any non-negative solution c of $\mathcal{A}(c) = b^\delta$ contains:

$$Z_\delta = \{u \in I^d; \delta'(\max(u)) = 0 \text{ or } \exists i \text{ such that } u_i < \max(u) \text{ and } \delta'(u_i) = d\}. \tag{2.14}$$

Proposition 2.8. *Let $\delta \in \mathcal{D}$. If c is feasible then $c = 0$ a.e. on Z_δ (that is $c\mathbf{1}_{Z_\delta} = 0$ a.e.).*

Proof. Recall that $0 \le \delta' \le d$. Since $c \in L^1(I^d)$, the condition $\mathcal{A}_{d+1}(c) = b_{d+1}$, that is for all $r \in I$

$$\int_{I^d} c(u)\mathbf{1}_{\{\max(u)\le r\}}\,du = \int_0^r \delta'(s)\,ds,$$

implies, by the monotone class theorem, that for all measurable subsets H of I, we have:

$$\int_{I^d} c(u)\mathbf{1}_H(\max(u))\,du = \int_H \delta'(s)\,ds.$$

Since $c \ge 0$ a.e., we deduce that a.e. $c(u)\mathbf{1}_{\{\delta'(\max(u))=0\}} = 0$.

Next, notice that for all $r \in I$:

$$\begin{aligned}\int_{I^d} c(u)\left(\sum_{i=1}^d \mathbf{1}_{\{u_i<\max(u),u_i\le r\}}\right)du &= \sum_{i=1}^d\left(\int_{I^d} c(u)\mathbf{1}_{\{u_i\le r\}}\,du - \int_{I^d} c(u)\mathbf{1}_{\{u_i=\max(u),\max(u)\le r\}}\,du\right)\\ &= dr - \delta(r)\\ &= \int_0^r (d-\delta'(s))\,ds.\end{aligned}$$

This implies that a.e. $c(u)\left(\sum_{i=1}^d \mathbf{1}_{\{u_i<\max(u),\delta'(u_i)=d\}}\right) = 0$, that is:

$$c(u)\mathbf{1}_{\{\exists i \text{ such that } u_i<\max(u),\delta'(u_i)=d\}} = 0.$$

This gives the result. □

We define μ to be the Lebesgue measure restricted to $Z_\delta^c = I^d \setminus Z_\delta$: $\mu(du) = \mathbf{1}_{Z_\delta^c}(u)du$. We define, for $f \in L^1(I^d,\mu)$:

$$\mathcal{I}^\mu(f) = \int_{I^d} f(u)\log(f(u))\,\mu(du).$$

From Proposition 2.8 we can deduce that if c is feasible then $\mathcal{I}^\mu(c) = \mathcal{I}_d(c)$. Let us also define, for $1 \le i \le d$, $r \in I$:

$$\mathcal{A}_i^\mu(c)(r) = \int_{I^d} c(u)\mathbf{1}_{\{u_i\le r\}}\,\mu(du), \quad \text{and} \quad \mathcal{A}_{d+1}^\mu(c)(r) = \int_{I^d} c(u)\mathbf{1}_{\{\max(u)\le r\}}\,\mu(du).$$

The corresponding optimization problem (P_μ^δ) is given by :

$$\text{minimize } \mathcal{I}^\mu(c) \text{ subject to } \begin{cases} \mathcal{A}^\mu(c) = b^\delta, \\ c \ge 0\ \mu\text{-a.e. and } c \in L^1(I^d,\mu), \end{cases} \tag{P_μ^δ}$$

with $\mathcal{A}^\mu = (\mathcal{A}_i^\mu, 1 \le i \le d+1)$. For $f \in L^1(I^d,\mu)$, we define:

$$f^\mu = \begin{cases} f \text{ on } Z_\delta^c, \\ 0 \text{ on } Z_\delta. \end{cases}$$

Using Proposition 2.8, we easily get the following Corollary.

Corollary 2.9. *If c is a solution of (P_μ^δ), then c^μ is a solution of (P^δ). If c is a solution of (P^δ), then it is also a solution of (P_μ^δ).*

2.5 Proof of Proposition 2.4

2.5.1 Form of the optimal solution

Let $(\mathcal{A}^\mu)^* : L^\infty(I)^{d+1} \to L^\infty(I^d, \mu)$ be the adjoint of $\mathcal{A}^\mu$. We will use Theorem 2.9. from [25] on abstract entropy minimization, which we recall here, adapted to the context of (P_μ^δ).

Theorem 2.10 (Borwein, Lewis and Nussbaum [25])**.** *Suppose there exists $c > 0$ μ-a.e. which is feasible for (P_μ^δ). Then there exists a unique optimal solution, c^*, to (P_μ^δ). Furthermore, we have $c^* > 0$ μ-a.e. and there exists a sequence $(\lambda^n, n \in \mathbb{N})$ of elements of $L^\infty(I)^{d+1}$ such that:*

$$\int_{I^d} c^*(x) \left| (\mathcal{A}^\mu)^*(\lambda^n)(x) - \log(c^*(x)) \right| \mu(dx) \xrightarrow[n\to\infty]{} 0. \tag{2.15}$$

We first compute $(\mathcal{A}^\mu)^*$. For $\lambda = (\lambda_i, 1 \le i \le d+1) \in L^\infty(I)^{d+1}$ and $f \in L^1(I^d, \mu)$, we have:

$$\begin{aligned}
\langle (\mathcal{A}^\mu)^*(\lambda), f \rangle &= \langle \lambda, \mathcal{A}^\mu(f) \rangle \\
&= \sum_{i=1}^d \int_I \lambda_i(r) \int_{I^d} f(x) \mathbf{1}_{\{x_i \le r\}} d\mu(x) dr + \int_I \lambda_{d+1}(r) \int_{I^d} f(x) \mathbf{1}_{\{\max(x) \le r\}} d\mu(x) dr \\
&= \int_{I^d} f(x) \left(\sum_{i=1}^d \Lambda_i(x_i) + \Lambda_{d+1}(\max(x)) \right) d\mu(x),
\end{aligned}$$

where we used the definition of the adjoint operator for the first equality, Fubini's theorem for the second, and the following notation for the third equality:

$$\Lambda_i(x_i) = \int_I \lambda_i(r) \mathbf{1}_{\{r \ge x_i\}} \, dr, \quad \text{and} \quad \Lambda_{d+1}(t) = \int_I \lambda_{d+1}(r) \mathbf{1}_{\{r \ge t\}} \, dr.$$

Thus, we can set for $\lambda \in L^\infty(I)^{d+1}$ and $x \in I^d$:

$$(\mathcal{A}^\mu)^*(\lambda)(x) = \sum_{i=1}^d \Lambda_i(x_i) + \Lambda_{d+1}(\max(x)). \tag{2.16}$$

Now we are ready to prove that the optimal solution c^* of (P_μ^δ) is the product of measurable univariate functions.

Lemma 2.11. *Let $\delta \in \mathcal{D}_0$ such that $\Sigma_\delta = \{0, 1\}$. Suppose that there exists $c > 0$ μ-a.e.which is feasible for (P_μ^δ). Then there exist a^*, b^* non-negative, measurable functions defined on I such that*

$$c^*(u) = b^*(\max(u)) \prod_{u_i \ne \max(u)} a^*(u_i) \quad \mu\text{-a.e.}$$

with $a^*(s) = 0$ if $\delta'(s) = d$ and $b^*(s) = 0$ if $\delta'(s) = 0$.

Proof. According to Theorem 2.10, there exists a sequence $(\lambda^n, n \in \mathbb{N})$ of elements of $L^\infty(I)^{d+1}$ such that the optimal solution, say c^*, satisfies (2.15). This implies, thanks to (2.16), that there exist $d+1$ sequences $(\Lambda_i^n, n \in \mathbb{N}, 1 \le i \le d+1)$ of elements of $L^\infty(I)$ such that the following convergence holds in $L^1(I^d, c^*\mu)$:

$$\sum_{i=1}^d \Lambda_i^n(u_i) + \Lambda_{d+1}^n(\max(u)) \xrightarrow[n\to\infty]{} \log(c^*(u)). \tag{2.17}$$

Arguing as in Proposition 2.7 and since Z_δ^c, the support of μ, is symmetric, we deduce that c^* is symmetric. Therefore we shall only consider functions supported on the set $\triangle = \{u \in I^d; u_d = \max(u)\}$. The convergence (2.17) holds in $L^1(\triangle, c^*\mu)$. For simplicity, we introduce the

functions $\Gamma_i^n \in L^\infty(I)$ defined by $\Gamma_i^n = \Lambda_i^n$ for $1 \le i \le d-1$, and $\Gamma_d^n = \Lambda_d^n + \Lambda_{d+1}^n$. Then we have in $L^1(\triangle, c^*\mu)$:

$$\sum_{i=1}^{d} \Gamma_i^n(u_i) \xrightarrow[n\to\infty]{} \log(c^*(u)). \tag{2.18}$$

We first assume that there exist Γ_i, $1 \le i \le d$ measurable functions defined on I such that μ-a.e. on $\triangle$:

$$\sum_{i=1}^{d} \Gamma_i(u_i) = \log(c^*(u)). \tag{2.19}$$

The symmetric property of $c^*(u)$ seen in Proposition 2.7 implies we can choose $\Gamma_i = \Gamma$ for $1 \le i \le d-1$ up to adding a constant to Γ_d. Set $a^* = \exp(\Gamma)$ and $b^* = \exp(\Gamma_d)$ so that μ-a.e. on $\triangle$:

$$c^*(u) = b^*(u_d) \prod_{i=1}^{d-1} a^*(u_i). \tag{2.20}$$

Recall $\mu(du) = \mathbf{1}_{Z_\delta^c}(u)\, du$. From the definition (2.14) of Z_δ, we deduce that without loss of generality, we can assume that $a^*(u_i) = 0$ if $\delta'(u_i) = d$ and $b^*(u_d) = 0$ if $\delta'(u_d) = 0$. Use the symmetry of c^* to conclude.

To complete the proof, we now show that (2.19) holds for Γ and Γ_d measurable functions. We introduce the notation $u_{(-i)} = (u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_d) \in I^{d-1}$. Let us define the probability measure $P(dx) = c^*(x)\mathbf{1}_\triangle(x)\mu(dx)/\int_\triangle c^*(y)\mu(dy)$ on I^d. We fix j, $1 \le j \le d-1$. In order to apply Proposition 2 of [153], which would ensure the existence of the limiting functions Γ_i, $1 \le i \le d$, we first check that P is absolutely continuous with respect to $P_1^j \otimes P_2^j$, where $P_1^j(du_{(-j)}) = \int_{u_j \in I} P(du_{(-j)}du_j)$ and $P_2^j(du_j) = \int_{u_{(-j)} \in I^{d-1}} P(du_{(-j)}du_j)$ are the marginals of P. Notice the following equivalence of measures:

$$P(du) \sim \mathbf{1}_\triangle(u) \prod_{i=1}^{d-1} \mathbf{1}_{\{\delta'(u_i)\neq d\}} \mathbf{1}_{\{\delta'(u_d)\neq 0\}}\, du. \tag{2.21}$$

Let $B \subset I^{d-1}$ be measurable. We have:

$$P_1(B) = 0 \iff \int_{I^d} \mathbf{1}_\triangle(u) \prod_{i=1}^{d-1} \mathbf{1}_{\{\delta'(u_i)\neq d\}} \mathbf{1}_{\{\delta'(u_d)\neq 0\}} \mathbf{1}_B(u_{(-j)})\, du = 0.$$

By Fubini's theorem this last equality is equivalent to:

$$\int_{I^{d-1}} \prod_{\substack{i=1\\ i\neq j}}^{d-1} \left(\mathbf{1}_{\{\delta'(u_i)\neq d\}} \mathbf{1}_{\{u_i \le u_d\}}\right) \mathbf{1}_{\{\delta'(u_d)\neq 0\}} \mathbf{1}_B(u_{(-j)}) \left(\int_I \mathbf{1}_{\{0\le u_j \le u_d\}} \mathbf{1}_{\{\delta'(u_j)\neq d\}}\, du_j\right) du_{(-j)} = 0. \tag{2.22}$$

Since for $\varepsilon > 0$, $\delta(\varepsilon) < \varepsilon < d\varepsilon$, we have $\int_I \mathbf{1}_{\{0\le u_j \le s\}} \mathbf{1}_{\{\delta'(u_j)\neq d\}}\, du_j > 0$ for all $s \in I$. Therefore (2.22) is equivalent to

$$\int_{I^{d-1}} \prod_{i=1, i\neq j}^{d-1} \left(\mathbf{1}_{\{\delta'(u_i)\neq d\}} \mathbf{1}_{\{u_i \le u_d\}}\right) \mathbf{1}_{\{\delta'(u_d)\neq 0\}} \mathbf{1}_B(u_{(-j)})\, du_{(-j)} = 0.$$

This implies that there exists $h > 0$ a.e. on I^{d-1} such that

$$P_1^j(du_{(-j)}) = h(u_{(-j)}) \prod_{i=1, i\neq j}^{d-1} \left(\mathbf{1}_{\{\delta'(u_i)\neq d\}} \mathbf{1}_{\{u_i \le u_d\}}\right) \mathbf{1}_{\{\delta'(u_d)\neq 0\}} du_{(-j)}.$$

Similarly we have for $B' \subset I$ that $P_2^j(B') = 0$ if and only if

$$\int_I \mathbf{1}_{\{\delta'(u_j)\neq d\}} \mathbf{1}_{B'}(u_j) \left(\int_{I^{d-1}} \prod_{i=1, i\neq j}^{d-1} \left(\mathbf{1}_{\{\delta'(u_i)\neq d\}} \mathbf{1}_{\{u_i \le u_d\}}\right) \mathbf{1}_{\{\delta'(u_d)\neq 0\}}\, \mathbf{1}_{\{u_d \ge u_j\}} du_{(-j)}\right) du_j = 0. \tag{2.23}$$

Since, for $\varepsilon > 0$, $\delta(1) - \delta(1-\varepsilon) > 1 - (1-\varepsilon) = \varepsilon > 0$, there exists $g > 0$ a.e. on I such that $P_2^j(du_j) = g(u_j)\mathbf{1}_{\{\delta'(u_j)\neq d\}}du_j$. Therefore by (2.21) we deduce that P is absolutely continuous with respect to $P_1^j \otimes P_2^j$. Then according to Proposition 2 of [153], (2.18) implies that there exist measurable functions Φ_j and Γ_j defined respectively on I^{d-1} and I, such that $c^*\mu$-a.e. on $\triangle$:

$$\log(c^*(u)) = \Phi_j(u_{(-j)}) + \Gamma_j(u_j).$$

As μ-a.e. $c^* > 0$, this equality holds μ-a.e. on $\triangle$. Since we have such a representation for every $1 \leq j \leq d-1$, we can easily verify that there exists a measurable function Γ_d defined on I such that $\log(c^*(u)) = \sum_{i=1}^d \Gamma_i(u_i)$ μ-a.e. on $\triangle$.

□

2.5.2 Calculation of the optimal solution

Now we prove that the optimal solution to (P^δ), if it exists, is indeed c_δ.

Proposition 2.12. *Let $\delta \in \mathcal{D}_0$ such that $\Sigma_\delta = \{0,1\}$. If there exists a feasible solution c to (P^δ) such that $c > 0$ μ-a.e., then the optimal solution c^* to (P^δ) is c_δ given by (2.4).*

Proof. In Lemma 2.11 we have already shown that if an optimal solution exists for (P^δ), then it is of the form $c^*(u) = b^*(\max(u)) \prod_{u_i \neq \max(u)} a^*(u_i)$. Here we will prove that the constraints of (P^δ) uniquely determine the functions a^* and b^* up to a multiplicative constant, giving $c^* = c_\delta$. We set for $r \in I$:

$$A^*(r) = \int_0^r a^*(s)\,ds$$

which take values in $[0, +\infty]$. From $\mathcal{A}_{d+1}(c^*) = b_{d+1}^\delta$, we have for $r \in I$:

$$\begin{aligned}
\delta(r) &= \int_{I^d} c^*(u)\mathbf{1}_{\{\max(u)\leq r\}}\,du \\
&= \int_{I^d} b^*(\max(u)) \prod_{u_i \neq \max(u)} a^*(u_i)\mathbf{1}_{\{\max(u)\leq r\}}\,du \\
&= d\int_I (A^*(s))^{d-1} b^*(s)\mathbf{1}_{\{s\leq r\}}\,ds. \qquad (2.24)
\end{aligned}$$

Taking the derivative with respect to r gives a.e. on I:

$$\delta'(r) = d(A^*(r))^{d-1} b^*(r). \qquad (2.25)$$

This implies that $A^*(r)$ is finite for all $r \in [0,1)$ and thus $A^*(0) = 0$. Similarly, using that $\mathcal{A}_1(c^*) = b_1^\delta$, we get that for $r \in I$:

$$\begin{aligned}
1 - r &= \int_{I^d} c^*(u)\mathbf{1}_{\{u_1\geq r\}}\,du \\
&= \int_{I^d} b^*(\max(u)) \prod_{u_i \neq \max(u)} a^*(u_i)\mathbf{1}_{\{u_1\geq r\}}\,du \\
&= \int_{I^d} \prod_{i=2}^d \left(a^*(u_i)\mathbf{1}_{\{u_i\leq u_1\}}\right) b^*(u_1)\mathbf{1}_{\{u_1\geq r\}}\,du \\
&\qquad + (d-1)\int_{I^d} a^*(u_1) \prod_{i=3}^d \left(a^*(u_i)\mathbf{1}_{\{u_i\leq u_2\}}\right) b^*(u_2)\mathbf{1}_{\{u_2\geq u_1\geq r\}}\,du \\
&= \int_I (A^*(s))^{d-1} b^*(s)\mathbf{1}_{\{s\geq r\}}\,ds \\
&\qquad + (d-1)\int_I (A^*(s))^{d-2} b^*(s)(A^*(s) - A^*(r))\mathbf{1}_{\{s\geq r\}}\,ds \\
&= d\int_I (A^*(s))^{d-1} b^*(s)\mathbf{1}_{\{s\geq r\}}\,ds - (d-1)A^*(r)\int_I (A^*(s))^{d-2} b^*(s)\mathbf{1}_{\{s\geq r\}}\,ds.
\end{aligned}$$

Using this and (2.24) we deduce that for $r \in I$:

$$h(r) = (d-1)A^*(r) \int_I (A^*(s))^{d-2} b^*(s) \mathbf{1}_{\{s \geq r\}} \, ds. \tag{2.26}$$

Since $r > \delta(r)$ on $(0,1)$, we have that A^* and $\int_I (A^*(s))^{d-2} b^*(s) \mathbf{1}_{\{s \geq r\}} \, ds$ are positive on $(0,1)$. Dividing (2.25) by (2.26) gives a.e. for $r \in I$:

$$\frac{d-1}{d} \frac{\delta'(r)}{h(r)} = \frac{(A^*(r))^{d-2} b(r)}{\int_I (A^*(r))^{d-2} b^*(s) \mathbf{1}_{\{r \leq s \leq 1\}} \, ds}.$$

We integrate both sides to get for $r \in I$:

$$\frac{d-1}{d} \left(\log \left(\frac{h(r)}{h(1/2)} \right) - \int_{1/2}^r \frac{1}{h(s)} \, ds \right) = \log \left(\frac{\int_I (A^*(s))^{d-2} b^*(s) \mathbf{1}_{\{r \leq s \leq 1\}} \, ds}{\int_I (A^*(s))^{d-2} b^*(s) \mathbf{1}_{\{1/2 \leq s \leq 1\}} \, ds} \right).$$

Notice that the choice for the lower bound $1/2$ of the integral was arbitrary, see Remark 2.1. Taking the exponential yields:

$$\alpha h^{(d-1)/d}(r) \, \mathrm{e}^{-F(r)} = \int_I (A^*(s))^{d-2} b^*(s) \mathbf{1}_{\{r \leq s \leq 1\}} \, ds, \tag{2.27}$$

for some positive constant α. From (2.26) and (2.27), we derive:

$$A^*(r) = \frac{1}{\alpha(d-1)} h^{1/d}(r) \, \mathrm{e}^{F(r)}. \tag{2.28}$$

This proves that the function A^* is uniquely determined up to a multiplicative constant and so is a^*. With the help of (2.25) and (2.28), we can express b^* as, for $r \in I$:

$$b^*(r) = \frac{\delta'(r)(\alpha(d-1))^{d-1}}{d} \, \mathrm{e}^{-(d-1)F(r)}. \tag{2.29}$$

The function b^* is also uniquely determined up to a multiplicative constant. Therefore (2.25) implies that there is a unique c^* of the form (2.20) which solves $\mathcal{A}(c) = b^\delta$. (Notice however that the functions a^* and b^* are defined up to a multiplicative constant.) Then according to Proposition 2.2 we get that c_δ defined by (2.20) with a and b defined by (2.5) solves $\mathcal{A}(c) = b^\delta$, implying that c^* is equal to c_δ. □

2.5.3 Proof of Proposition 2.4

Let $\delta \in \mathcal{D}_0$ such that $\Sigma_\delta = \{0,1\}$. By construction, we have μ-a.e. $c_\delta > 0$. According to Proposition 2.2 and Remark 2.3, if $\mathcal{J}(\delta) < +\infty$, the copula density c_δ is feasible for (P^δ). Therefore Proposition 2.12 implies that it is the optimal solution as well. When $\mathcal{J}(\delta) = +\infty$, we show that there exists no feasible solution to c_δ, see the supplementary material.

2.6 Proof of Theorem 2.5

We first state an elementary Lemma, whose proof is left to the reader. For f a function defined on I^d and $0 \leq s < t \leq 1$, we define $f^{s,t}$ by, for $u \in I^d$:

$$f^{s,t}(u) = (t-s) f(s\mathbf{1} + u(t-s)).$$

Lemma 2.13. *If c is the density of a copula C such that $\delta_C(s) = s$ and $\delta_C(t) = t$ for some fixed $0 \leq s < t \leq 1$, then $c^{s,t}$ is also the density of a copula, and its diagonal section, $\delta^{s,t}$, is given by, for $r \in I$:*

$$\delta^{s,t}(r) = \frac{\delta_C(s + r(t-s)) - s}{t-s}.$$

According to Remark 2.6, it is enough to consider the case $\delta \in \mathcal{D}_0$, that is Σ_δ with zero Lebesgue measure. We shall assume that $\Sigma_\delta \neq \{0,1\}$. Since δ is continuous, we get that $I \setminus \Sigma_\delta$ can be written as the union of non-empty open disjoint intervals $((\alpha_j, \beta_j), j \in J)$, with $\alpha_j < \beta_j$ and J non-empty and at most countable. Set $\Delta_j = \beta_j - \alpha_j$. Since Σ_δ is of zero Lebesgue measure, we have $\sum_{j\in J} \Delta_j = 1$. We define also $S = \bigcup_{j\in J} [\alpha_j, \beta_j]^d$

For $s \in \Sigma_\delta$, notice that any feasible function c of (P^δ) satisfies for all $1 \leq i \leq d$:

$$\int_{I^d} c(u)\mathbf{1}_{\{u_i<s\}}\mathbf{1}_{D_i^c}(u)\,du = \int_{I^d} c(u)\mathbf{1}_{\{u_i<s\}}\,du - \int_{I^d} c(u)\mathbf{1}_{\{\max(u)<s\}}\,du = s - \delta(s) = 0,$$

where $D_i = \{u \in I^d \text{ such that } \forall j \neq i : u_j < s\}$. This implies that $c = 0$ a.e. on $I^d \setminus S$. We set $c^j = c^{\alpha_j,\beta_j}$ for $j \in J$. We deduce that if c is feasible for (P^δ), then we have that a.e.:

$$c(u) = \sum_{j\in J} \frac{1}{\Delta_j} c^j\left(\frac{u - \alpha_j\mathbf{1}}{\Delta_j}\right)\mathbf{1}_{(\alpha_j,\beta_j)^d}(u), \tag{2.30}$$

and:

$$\mathcal{I}_d(c) = \sum_{j\in J} \Delta_j\left(\mathcal{I}_d(c^j) - \log(\Delta_j)\right). \tag{2.31}$$

Thanks to Lemma 2.13, the condition $\mathcal{A}(c) = b^\delta$ is equivalent to $\mathcal{A}(c^j) = b^{\delta^j}$ for all $j \in J$. We deduce that the optimal solution of (P^δ), if it exists, is given by (2.30), where the functions c^j are the optimal solutions of (P^{δ^j}) for $j \in J$. Notice that by construction $\Sigma_{\delta^j} = \{0,1\}$. Thanks to Proposition 2.4, the optimal solution to (P^{δ^j}) exists if and only if we have $\mathcal{J}(\delta^j) < +\infty$; and if it exists it is given by c_{δ^j}. Therefore, if there exists an optimal solution to (P^δ), then it is c_δ given by (2.9). To conclude, we have to compute $\mathcal{I}_d(c_\delta)$. Recall that $x\log(x) \geq -1/\mathrm{e}$ for $x > 0$. We have:

$$\begin{aligned}
\mathcal{I}_d(c_\delta) &= \lim_{\varepsilon\downarrow 0} \sum_{j\in J} \Delta_j\left(\mathcal{I}_d(c^j) - \log(\Delta_j)\right)\mathbf{1}_{\{\Delta_j>\varepsilon\}} \\
&= \lim_{\varepsilon\downarrow 0} \sum_{j\in J} \Delta_j\left((d-1)\mathcal{J}(\delta^j) - \log(\Delta_j)\right)\mathbf{1}_{\{\Delta_j>\varepsilon\}} + \sum_{j\in J}\Delta_j\mathcal{G}(\delta^j) \\
&= \sum_{j\in J} \Delta_j\left((d-1)\mathcal{J}(\delta^j) - \log(\Delta_j)\right) + \sum_{j\in J}\Delta_j\mathcal{G}(\delta^j),
\end{aligned}$$

where we used the monotone convergence theorem for the first equality, (2.7) for the second and the fact that $\mathcal{G}(\delta)$ is uniformly bounded over $\mathcal{D}_0$ and the monotone convergence theorem for the last. Elementary computations yields:

$$(d-1)\mathcal{J}(\delta) = \sum_{j\in J} \Delta_j\left((d-1)\mathcal{J}(\delta^j) - \log(\Delta_j)\right) \quad \text{and} \quad \mathcal{G}(\delta) = \sum_{j\in J}\Delta_j\mathcal{G}(\delta^j).$$

So, we get:

$$\mathcal{I}_d(c_\delta) = (d-1)\mathcal{J}(\delta) + \mathcal{G}(\delta).$$

Since $\mathcal{G}(\delta)$ is uniformly bounded over $\mathcal{D}_0$, we get that $\mathcal{I}_d(c_\delta)$ is finite if and only if $\mathcal{J}(\delta)$ is finite. To end the proof, recall the definition of $\mathcal{I}(C_\delta)$ to conclude that $\mathcal{I}(C_\delta) = (d-1)\mathcal{J}(\delta) + \mathcal{G}(\delta)$.

2.7 Examples for $d = 2$

In this section we compute the density of the maximum entropy copula for various diagonal sections of popular bivariate copula families. In this Section, u and v will denote elements of I. The density for $d = 2$ is of the form $c_\delta(u,v) = a(\min(u,v))b(\max(u,v))$. For $(u,v) \in \triangle = \{(u,v) \in I^2, u \leq v\}$, the formula reads:

$$c_\delta(u,v) = \frac{\delta'(u)}{2\sqrt{h(u)}}\frac{2-\delta'(v)}{2\sqrt{h(v)}}\,\mathrm{e}^{-(F(v)-F(u))},$$

with h, F defined in (2.6). We illustrate these densities by displaying their isodensity lines or contour plots, and their diagonal cross-section φ defined as $\varphi(t) = c(t,t)$, $t \in I$.

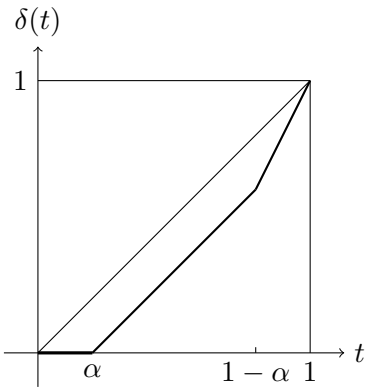


Figure 2.1 – Piecewise linear diagonal section (Section 2.7.1). Graph of δ with $\alpha = 0.2$.

2.7.1 Maximum entropy copula for a piecewise linear diagonal section

Let $\alpha \in (0, 1/2]$. Let us calculate the density of the maximum entropy copula in the case of the following diagonal section:

$$\delta(r) = (r-\alpha)\mathbf{1}_{(\alpha,1-\alpha)}(r) + (2r-1)\mathbf{1}_{[1-\alpha,1]}(r).$$

This example was considered for example in [133]. The limiting cases $\alpha = 0$ and $\alpha = 1/2$ correspond to the Fréchet-Hoeffding upper and lower bound copulas, respectively. However for $\alpha = 0$, $\Sigma_\delta = I$, therefore every copula C with this diagonal section gives $\mathcal{I}(C) = +\infty$. (In fact the only copula that has this diagonal section is the Fréchet-Hoeffding upper bound M defined by $M(u,v) = \min(u,v)$, $u, v \in I$.) When $\alpha \in (0, 1/2]$, $\mathcal{J}(\delta) < +\infty$ is satisfied, therefore we can apply Proposition 2.4 to compute the density of the maximum entropy copula. The graph of δ can be seen in Figure 2.1 for $\alpha = 0.2$. We compute the functions F, a and b:

$$F(r) = \begin{cases} \frac{1}{2}\log(\frac{r}{\alpha}) - \frac{1}{4\alpha} + \frac{1}{2} & \text{if } r \in [0,\alpha), \\ \frac{r}{2\alpha} - \frac{1}{4\alpha} & \text{if } t \in [\alpha, 1-\alpha), \\ \frac{1}{2}\log\left(\frac{\alpha}{1-r}\right) + \frac{1}{4\alpha} - \frac{1}{2} & \text{if } t \in [1-\alpha, 1], \end{cases}$$

$$a(r) = \frac{1}{\sqrt{\alpha}}\,\mathrm{e}^{-\frac{1}{4\alpha}+\frac{1}{2}}\,\mathbf{1}_{[0,\alpha]}(r) + \frac{1}{2\sqrt{\alpha}}\,\mathrm{e}^{\frac{r}{2\alpha}-\frac{1}{4\alpha}}\,\mathbf{1}_{(\alpha,1-\alpha)}(r),$$

and:

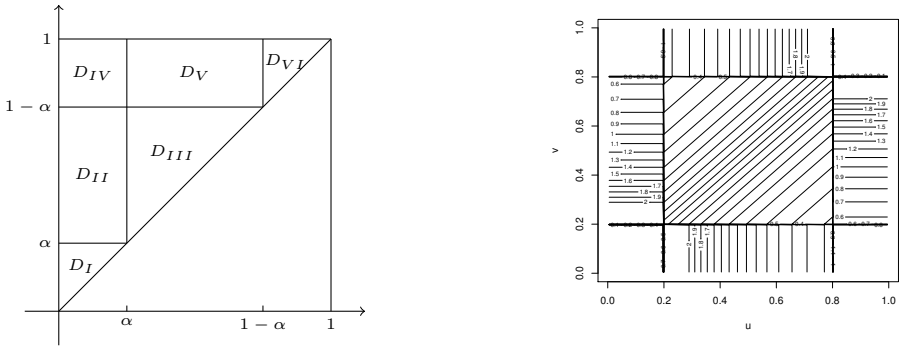
$$b(r) = \frac{1}{2\sqrt{\alpha}}\,\mathrm{e}^{-\frac{r}{2\alpha}+\frac{1}{4\alpha}}\,\mathbf{1}_{(\alpha,1-\alpha)}(r) + \frac{1}{\sqrt{\alpha}}\,\mathrm{e}^{-\frac{1}{4\alpha}+\frac{1}{2}}\,\mathbf{1}_{[1-\alpha,1]}(r).$$

The density $c_\delta(u,v)$ consists of six distinct regions on $\triangle$ as shown in Figure 2.2a and takes the values:

$$\begin{aligned} c_\delta(u,v) = \frac{1}{2\alpha}\,\mathrm{e}^{\frac{\alpha-v}{2\alpha}}\,\mathbf{1}_{\{(u,v)\in D_{II}\}} + \frac{1}{4\alpha}\,\mathrm{e}^{\frac{u-v}{2\alpha}}\,\mathbf{1}_{\{(u,v)\in D_{III}\}} \\ + \frac{1}{\alpha}\,\mathrm{e}^{\frac{2\alpha-1}{2\alpha}}\,\mathbf{1}_{\{(u,v)\in D_{IV}\}} + \frac{1}{2\alpha}\,\mathrm{e}^{\frac{u+\alpha-1}{2\alpha}}\,\mathbf{1}_{\{(u,v)\in D_{V}\}} \end{aligned} \tag{2.32}$$

Figure 2.2b shows the isodensity lines of c_δ. In the limiting case of $\alpha = \frac{1}{2}$, the diagonal section is given by $\delta(t) = \max(0, 2t-1)$,which is the pointwise lower bound for all elements in $\mathcal{D}$. Accordingly, it is the diagonal section of the Fréchet-Hoeffding lower bound copula W given by $W(u,v) = \max(0, u+v-1)$ for $u, v \in I$. All copulas having this diagonal section are of the following form:

$$D_{C_1,C_2}(u,v) = \begin{cases} W(u,v) & \text{if } (u,v) \in [0,1/2]^2 \cup [1/2,1]^2, \\ \frac{1}{2}C_1(2u, 2v-1) & \text{if } (u,v) \in [0,1/2] \times [1/2,1], \\ \frac{1}{2}C_2(2u-1, 2v) & \text{if } (u,v) \in [1/2,1] \times [0,1/2], \end{cases}$$



(a) Partition for c_δ (b) Isodensity lines of c_δ

Figure 2.2 – Piecewise linear diagonal section (Section 2.7.1). The partition and the isodensity lines of c_δ.

where C_1 and C_2 are copula functions. Recall that the independent copula Π with uniform density $c_\Pi = 1$ on I^2 minimizes $\mathcal{I}(C)$ over $\mathcal{C}$. According to (2.32), the maximum entropy copula with diagonal section δ is $D_{\Pi,\Pi}$. This corresponds to choosing the maximum entropy copulas on $[0, 1/2] \times [1/2, 1]$ and $[1/2, 1] \times [0, 1/2]$.

2.7.2 Maximum entropy copula for $\delta(t) = t^\alpha$

Let $\alpha \in (1, 2]$. We consider the family of diagonal sections given by $\delta(t) = t^\alpha$. This corresponds to the Gumbel family of copulas and also to the family of Cuadras-Augé copulas. The Gumbel copula with parameter $\theta \in [1, \infty)$ is an Archimedean copula defined as, for $u, v \in I$:

$$C^G(u, v) = \varphi_\theta^{-1}(\varphi_\theta(u) + \varphi_\theta(v))$$

with generator function $\varphi_\theta(t) = (-\log(t))^\theta$. Its diagonal section is given by $\delta^G(t) = t^{2^{\frac{1}{\theta}}} = t^\alpha$ with $\alpha = 2^{\frac{1}{\theta}}$. The Cuadras-Augé copula with parameter $\gamma \in (0, 1)$ is defined as, for $u, v \in I$:

$$C^{CA}(u, v) = \min(uv^{1-\gamma}, u^{1-\gamma}v).$$

It is a subclass of the two parameter Marshall-Olkin family of copulas given by:

$$C^M(u, v) = \min(u^{1-\gamma_1}v, uv^{1-\gamma_2}).$$

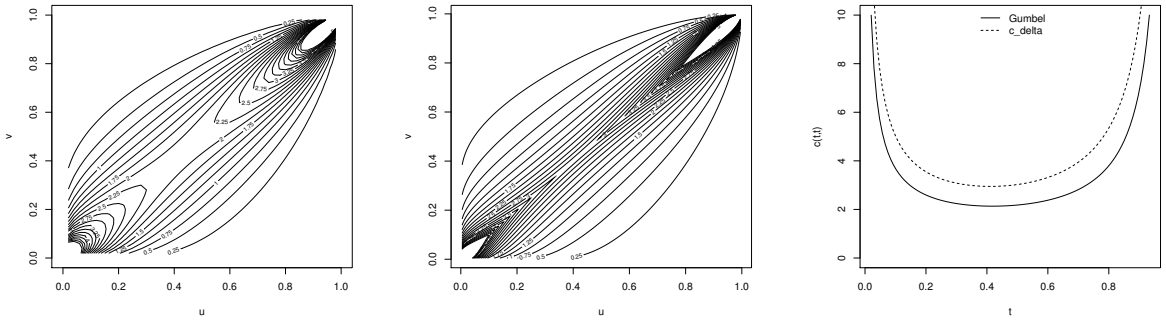
The diagonal section of C^{CA} is given by $\delta(t) = t^{2-\gamma} = t^\alpha$ with $\alpha = 2 - \gamma$. While the Gumbel copula is absolutely continuous, the Cuadras-Augé copula is not, although it has full support. Since $\mathcal{J}(\delta) < +\infty$, we can apply Proposition 2.4. To give the density of the maximum entropy copula, we have to calculate $F(v) - F(u)$. Elementary computations yield:

$$F(v) - F(u) = \frac{1}{2}\int_u^v \frac{ds}{s - s^\alpha} = \frac{1}{2}\log\left(\frac{v}{u}\right) - \frac{1}{2\alpha - 2}\log\left(\frac{1 - v^{\alpha-1}}{1 - u^{\alpha-1}}\right).$$

The density c_δ is therefore given by, for $(u, v) \in \triangle$:

$$c_\delta(u, v) = \frac{\alpha}{4}\,\frac{2 - \alpha u^{\alpha-1}}{(1 - u^{\alpha-1})^{\alpha/(2\alpha-2)}}\, v^{\alpha-2}(1 - v^{\alpha-1})^{(2-\alpha)/(2\alpha-2)}.$$

Figure 2.3 represents the isodensity lines of the Gumbel and the maximum entropy copula c_δ with common parameter $\alpha = 2^{\frac{1}{3}}$, which corresponds to $\theta = 3$ for the Gumbel copula. We have also added a graph of the diagonal cross-section of the two densities. In the limiting case of $\alpha = 2$, the above formula gives $c_\delta(u, v) = 1$, which is the density of the independent copula Π, which is also maximizes the entropy on the entire set of copulas.



(a) Gumbel (b) c_δ (c) Diagonal cross-section

Figure 2.3 – Power function diagonal section (Section 2.7.2). Isodensity lines and the diagonal cross-section of copulas with diagonal section $\delta(t) = t^\alpha$, $\alpha = 2^{\frac{1}{3}}$.

2.7.3 Maximum entropy copula for the Farlie-Gumbel-Morgenstern diagonal section

Let $\theta \in [-1, 1]$. The Farlie-Gumbel-Morgenstern family of copulas (FGM copulas for short) are defined as:

$$C(u, v) = uv + \theta uv(1 - u)(1 - v).$$

These copulas are absolutely continuous with densities $c(u, v) = 1+\theta(1-2u)(1-2v)$. Its diagonal section δ_θ is given by:

$$\delta_\theta(t) = t^2 + \theta t^2(1 - t)^2 = \theta t^4 - 2\theta t^3 + (1 + \theta)t^2.$$

Since $\delta_\theta(t) < t$ on $(0, 1)$ and it verifies $\mathcal{J}(\delta) < +\infty$, we can apply Proposition 2.4 to calculate the density of the maximum entropy copula. For $F(r)$, we have:

$$F(r) = \begin{cases} \frac{1}{2}\log\left(\frac{r}{1-r}\right) + \frac{\theta}{\sqrt{4\theta-\theta^2}}\arctan\left(\frac{2\theta r-\theta}{\sqrt{4\theta-\theta^2}}\right) & \text{if } \theta \in (0, 1], \\ \frac{1}{2}\log\left(\frac{r}{1-r}\right) & \text{if } \theta = 0, \\ \frac{1}{2}\log\left(\frac{r}{1-r}\right) - \frac{\theta}{\sqrt{\theta^2-4\theta}}\operatorname{arctanh}\left(\frac{2\theta r-\theta}{\sqrt{\theta^2-4\theta}}\right) & \text{if } \theta \in [-1, 0). \end{cases}$$

The density c_δ is given by, for $\theta \in (0, 1]$ and $(u, v) \in \triangle$:

$$c_\delta(u, v) = \frac{(1 - 2\theta u^3 + 3\theta u^2 + (1 + \theta)u)}{(1 - u)\sqrt{\theta u^2 - \theta u + 1}} \frac{(2\theta v^2 + 3\theta v + (1 + \theta))}{\sqrt{\theta v^2 - \theta v + 1}} \exp\left(-\frac{\theta}{\sqrt{4\theta - \theta^2}}\left(\arctan\left(\frac{2\theta v - \theta}{\sqrt{4\theta - \theta^2}}\right) - \arctan\left(\frac{2\theta u - \theta}{\sqrt{4\theta - \theta^2}}\right)\right)\right).$$

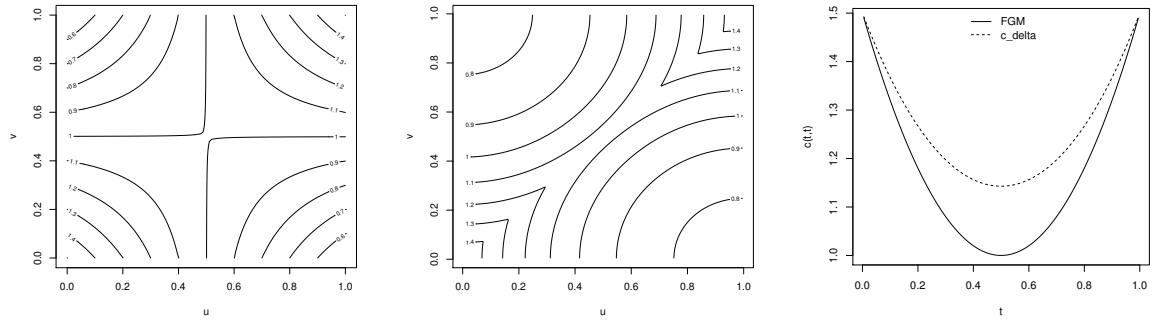
Figure 2.4 illustrates the isodensities of the FGM copula and the maximum entropy copula with the same diagonal section for $\theta = 0.5$ as well as the diagonal cross-section of their densities.

The case of $\theta = 0$ corresponds once again to the diagonal section $\delta(t) = t^2$, and the formula gives the density of the independent copula Π, accordingly.

2.7.4 Maximum entropy copula for the Ali-Mikhail-Haq diagonal section

Let $\theta \in [-1, 1]$. The Ali-Mikhail-Haq (AMH for short) family of copulas are defined as:

$$C(u, v) = \frac{uv}{1 - \theta(1 - u)(1 - v)}.$$



(a) FGM (b) c_δ (c) Diagonal cross-section

Figure 2.4 – FGM diagonal section (Section 2.7.3). Isodensity lines and the diagonal cross-section of copulas with diagonal section $\delta(t) = \theta t^4 - 2\theta t^3 + (1+\theta)t^2$, $\theta = 0.5$.

This is a family of absolutely continuous copulas whose diagonal section is given by:

$$\delta(t) = \frac{t^2}{1-\theta(1-t)^2}.$$

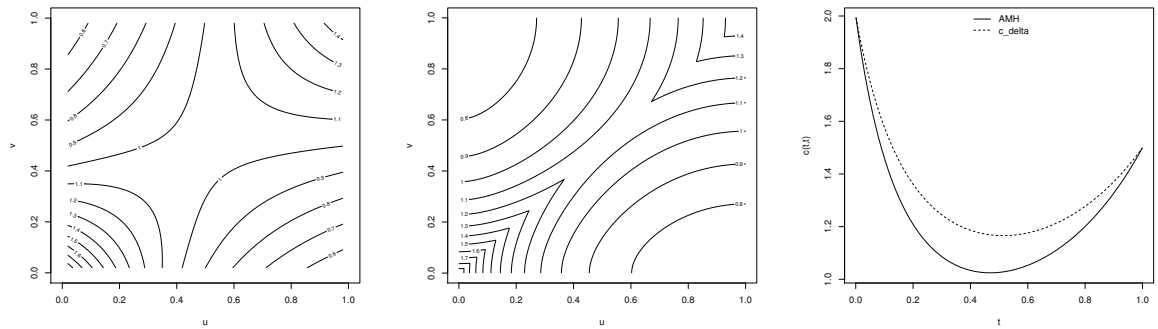
Once again, $\delta_\theta(t) < t$ on $(0,1)$ and $\mathcal{J}(\delta) < +\infty$ is verified, so we can apply Proposition 2.4 to calculate the density of the maximum entropy copula. For $0 \leq u \leq v \leq 1$:

$$F(v) - F(u) = \frac{1}{2}\left(\ln\left(\frac{v}{u}\right) - \ln\left(\frac{1-v}{1-u}\right) + \ln\left(\frac{\theta v + 1 - \theta}{\theta u + 1 - \theta}\right)\right).$$

Then c_δ is given by, for $(u,v) \in \triangle$:

$$c_\delta(u,v) = \frac{1 + \theta u - 2\theta(1-u) + \theta^2(1-u)^3}{(1-\theta(1-u)^2)^{\frac{3}{2}}}\left(1-\theta(1-v)^2\right)^{-\frac{3}{2}}.$$

In the case of $\theta = 0$, the AMH copula reduces to the independent copula Π. We illustrate the density of the AMH copula and the corresponding maximum entropy copula with $\theta = 0.5$ in Figure 2.5 and $\theta = -0.5$ in Figure 2.6.



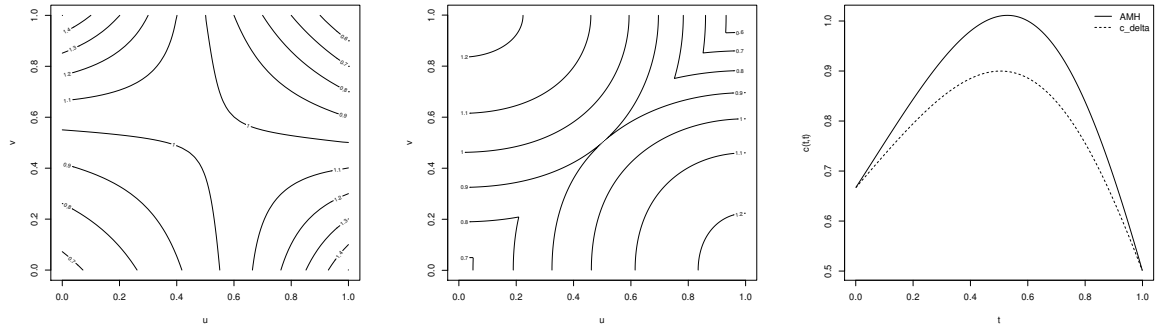
(a) AMH (b) c_δ (c) Diagonal cross-section

Figure 2.5 – AMH diagonal section (Section 2.7.4). Isodensity lines and the diagonal cross-section of copulas with diagonal section $\delta(t) = \frac{t^2}{1-\theta(1-t)^2}$, $\theta = 0.5$.

2.7.5 Maximum entropy copula for the Gaussian diagonal section

The Gaussian (normal) copula takes the form:

$$C_\rho(u,v) = \Phi_\rho\left(\Phi^{-1}(u), \Phi^{-1}(v)\right),$$



(a) AMH (b) c_δ (c) Diagonal cross-section

Figure 2.6 – AMH diagonal section (Section 2.7.4). Isodensity lines and the diagonal cross-section of copulas with diagonal section $\delta(t) = \frac{t^2}{1-\theta(1-t)^2}$, $\theta = -0.5$.

with Φ_ρ the joint cumulative distribution function of a two-dimensional normal random variable with standard normal marginals and correlation parameter $\rho \in [-1, 1]$, and Φ^{-1} the quantile function of the standard normal distribution. The density c_ρ of C_ρ can be written as:

$$c_\rho(u, v) = \frac{\varphi_\rho\left(\Phi^{-1}(u), \Phi^{-1}(v)\right)}{\varphi(\Phi^{-1}(u))\varphi(\Phi^{-1}(v))},$$

where φ and φ_ρ stand for respectively the densities of a standard normal distribution and a two-dimensional normal distribution with correlation parameter ρ, respectively. The diagonal section and its derivative are given by:

$$\delta_\rho(t) = \Phi_\rho\left(\Phi^{-1}(t), \Phi^{-1}(t)\right), \ \delta'_\rho(t) = 2\Phi\left(\sqrt{\frac{1-\rho}{1+\rho}}\Phi^{-1}(t)\right). \tag{2.33}$$

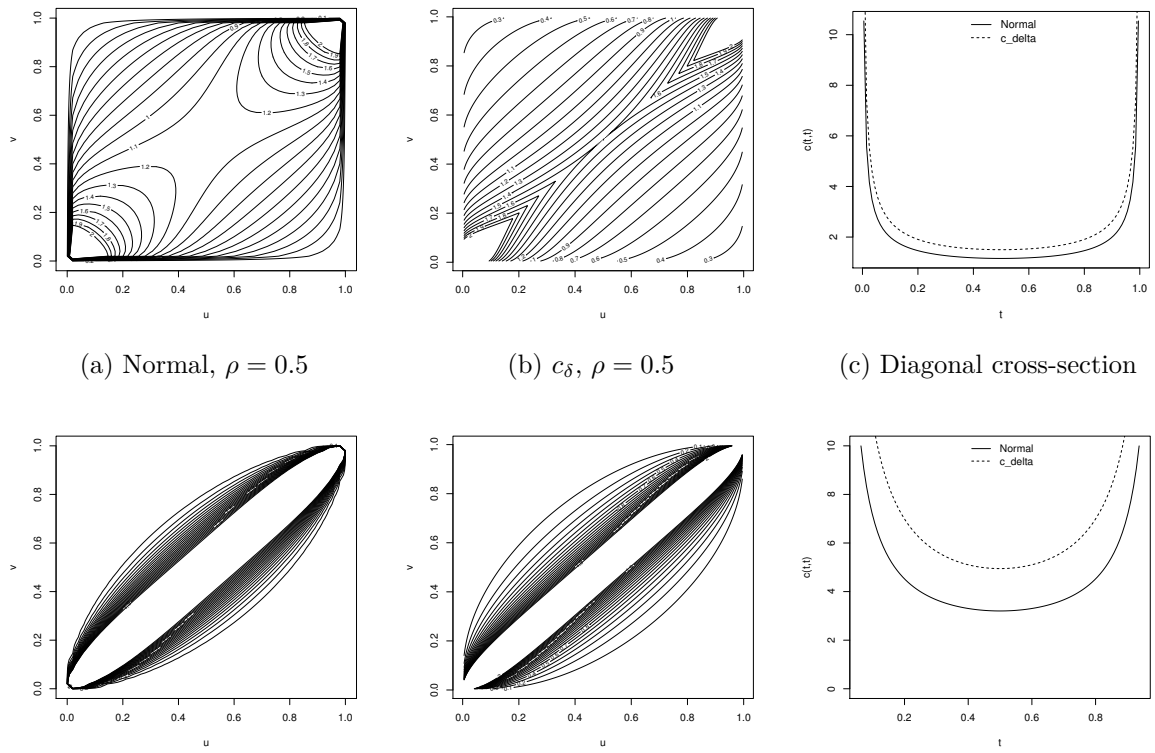
Since δ_ρ verifies $\delta_\rho(t) < t$ on $(0, 1)$ and $\mathcal{J}(\delta_\rho) < +\infty$, we can apply Proposition 2.4 to calculate the density of the maximum entropy copula. We have calculated numerically the density of the maximum entropy copula with diagonal section δ_ρ for $\rho = 0.95, 0.5, -0.5$ and -0.95. The comparison between these densities and the densities of the corresponding normal copula can be seen in Figures 2.7,2.8 and 2.9. In the limiting case when ρ goes up to 1, we observe a similar behaviour of c_ρ and c_{δ_ρ}, and we get the limiting diagonal $\delta(t) = t$ of the Fréchet-Hoeffding upper bound M given by $M(u, v) = \min(u, v)$, which does not have a density. We observe a very different behaviour of c_ρ and c_{δ_ρ} in the case of $\rho < 0$. In the limiting case when ρ goes down to -1, we get the diagonal $\delta(t) = \max(0, 2t - 1)$, which we have studied earlier in Section 2.7.1.

2.7.6 Comparison of conditional extreme event probabilities in the Gaussian case

We compare the conditional probabilities of extreme values of a pair of random variables (X_1, X_2) which has bivariate normal distribution with standard normal marginals and correlation coefficient ρ, with a pair of random variables (Y_1, Y_2) whose marginals are also standard normal, but has copula c_δ, where δ is the diagonal of the copula of (X_1, X_2). We compute the conditional probabilities $\mathbb{P}(X_1 \geq \alpha t | X_2 = t)$ and $\mathbb{P}(Y_1 \geq \alpha t | Y_2 = t)$ with $\alpha \geq 1$ and consider their asymptotic behaviour when t goes to infinity. This comparison is motivated by consideration of correlated defaults in mathematical finance, see Section 10.8 in [157]. (Notice however the parameters of upper tail dependence of the two copulas are the same since they have the same diagonal.)

Since by construction $\max(X_1, X_2)$ has the same distribution as $\max(Y_1, Y_2)$, and X_1, X_2, Y_1 and Y_2 have the same distribution, we deduce that $\min(X_1, X_2)$ has the same distribution as $\min(Y_1, Y_2)$. We deduce that for all $t \in \mathbb{R}$:

$$\mathbb{P}(X_1 \geq t | X_2 = t) = -\frac{\partial_t \mathbb{P}(\min(X_1, X_2) \geq t)}{\varphi(t)} = -\frac{\partial_t \mathbb{P}(\min(Y_1, Y_2) \geq t)}{\varphi(t)} = \mathbb{P}(Y_1 \geq t | Y_2 = t).$$



(a) Normal, $\rho = 0.5$ (b) c_δ, $\rho = 0.5$ (c) Diagonal cross-section

(d) Normal, $\rho = 0.95$ (e) c_δ, $\rho = 0.95$ (f) Diagonal cross-section

Figure 2.7 – Gaussian diagonal section (Section 2.7.5). Isodensity lines and the diagonal cross-section of copulas with diagonal section given by (2.33), with $\rho = 0.5$ and $\rho = 0.95$.

From now on, we shall consider $\alpha > 1$. For $k \in \mathbb{R}$, we recall the notations $h(t) = O(t^k)$ for t large which means that $\limsup_{t\to+\infty} t^{-k}|h(t)| < +\infty$, and $f(t) \ll g(t)$ for t large which means that f and g are positive for t large and $\limsup_{t\to\infty} f(t)/g(t) = 0$. The proof of the next Lemma is given in the Appendix.

Lemma 2.14. *Let $\alpha > 1$ and $\rho \in (-1, 1)$. We have for t large:*

$$\mathbb{P}(X_1 \geq \alpha t | X_2 = t) = \kappa_{\rho,\alpha} \mathbb{P}(Y_1 \geq \alpha t | Y_2 = t)\, \mathrm{e}^{-\Delta_{\rho,\alpha} t^2/2}(1 + O(t^{-2})), \tag{2.34}$$

with:

$$\kappa_{\rho,\alpha} = \frac{\alpha(1-\rho)}{(\alpha-\rho)} \quad \textit{and} \quad \Delta_{\rho,\alpha} = \frac{\rho(\alpha-1)}{1-\rho^2}\left((\alpha+1)\rho - 2\right).$$

We deduce that:

— for $\rho > 0$ and $\alpha > 2/\rho - 1$ or $\rho < 0$, we have $\Delta_{\rho,\alpha} > 0$ and thus $\mathbb{P}(X_1 \geq \alpha X_2 | X_2 = t) \ll \mathbb{P}(Y_1 \geq \alpha Y_2 | Y_2 = t)$ for t large,

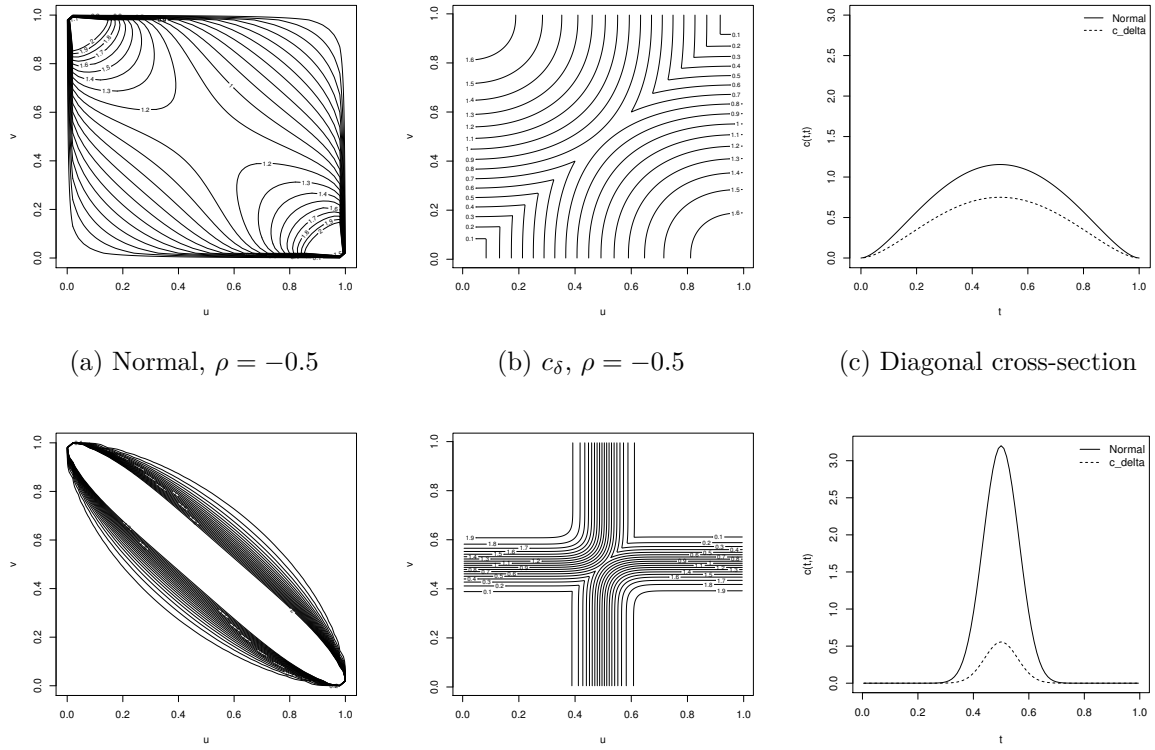
— for $\rho > 0$ and $1 < \alpha < 2/\rho - 1$, we have $\Delta_{\rho,\alpha} < 0$ and thus $\mathbb{P}(X_1 \geq \alpha X_2 | X_2 = t) \gg \mathbb{P}(Y_1 \geq \alpha Y_2 | Y_2 = t)$ for t large.

In conclusion, in the positive correlation case, the maximum entropy copula gives more weight to the extremal conditional probabilities for large values of α.

Remark 2.15. Similar computations as in the proof of Lemma 2.14 give that for $\rho > 0$, $\rho \leq \alpha < 1$:

$$\mathbb{P}\left(\alpha t \leq X_1 \leq t | X_2 = t\right) = \bar{\Phi}\left(\frac{\alpha-\rho}{\sqrt{1-\rho^2}} t\right)\left(1 + O(t^{-2})\right),$$

$$\mathbb{P}\left(\alpha t \leq Y_1 \leq t | Y_2 = t\right) = \bar{\Phi}\left(\alpha\sqrt{\frac{1-\rho}{1+\rho}} t\right)\left(1 + O(t^{-2})\right),$$



(a) Normal, $\rho = -0.5$ (b) c_δ, $\rho = -0.5$ (c) Diagonal cross-section

(d) Normal, $\rho = -0.95$ (e) c_δ, $\rho = -0.95$ (f) Diagonal cross-section

Figure 2.8 – Gaussian diagonal section (Section 2.7.5). Isodensity lines and the diagonal cross-section of copulas with diagonal section given by (2.33), with $\rho = -0.5$ and $\rho = -0.95$

with $\bar{\Phi} = 1 - \Phi$, the survival function of the standard Gaussian distribution. Using (2.42), we have $\mathbb{P}\left(\alpha t \le X_1 \le t | X_2 = t\right) \gg \mathbb{P}(Y_1 \ge \alpha Y_2 | Y_2 = t)$ for t large. This means that the maximum entropy copula gives less weight to the "non-worse" case, when the first variable takes also large values, but stays less than the second variable.

2.8 Appendix

2.8.1 Calculation of the entropy of C_δ

In this section, we show that (2.7) of Proposition 2.2 holds. Let us first introduce some notations. Let $\varepsilon \in (0, 1/2)$. Since $x \log(x) \ge -1/\mathrm{e}$ for $x > 0$, we deduce by the monotone convergence theorem that:

$$\mathcal{I}(C_\delta) = \lim_{\varepsilon \downarrow 0} \mathcal{I}_\varepsilon(C_\delta), \tag{2.35}$$

with:

$$\mathcal{I}_\varepsilon(C_\delta) = \int_{[\varepsilon, 1-\varepsilon]^d} c_\delta(x) \log(c_\delta(x))\, dx.$$

Using $\delta(t) \le t$ and that δ is a non-decreasing, d-Lipschitz function, we get that for $t \in I$:

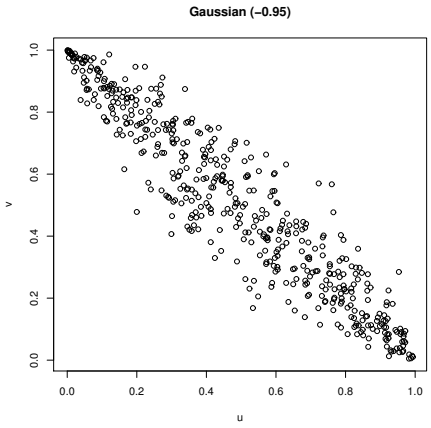
$$0 \le h(t) \le \min(t, (d-1)(1-t)) \le (d-1)\min(t, 1-t). \tag{2.36}$$

We set:

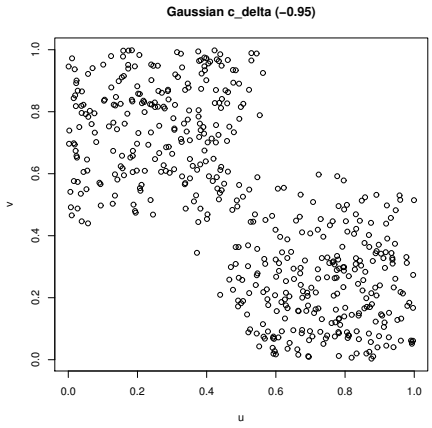
$$w(t) = a(t)\, \mathrm{e}^{-F(t)} = \frac{d - \delta'(t)}{d} h^{-1+1/d}(t). \tag{2.37}$$

From the symmetric property of c_δ, we have that

$$\mathcal{I}_\varepsilon(C_\delta) = J_1(\varepsilon) + J_2(\varepsilon) - J_3(\varepsilon), \tag{2.38}$$



(a) Normal, $\rho = -0.95$



(b) $C_\delta, \rho = -0.95$

Figure 2.9 – Gaussian diagonal section (Section 2.7.5). Sample of 500 drawn from the Gaussian copula with $\rho = -0.95$ and from the corresponding C_δ

with:

$$J_1(\varepsilon) = d \int_{[\varepsilon,1-\varepsilon]^d} c_\delta(x) \mathbf{1}_{\{\max(x)=x_d\}} \left(\sum_{i=1}^{d-1} \log\left(w(x_i)\right) \right) dx,$$
$$J_2(\varepsilon) = d \int_{[\varepsilon,1-\varepsilon]^d} c_\delta(x) \mathbf{1}_{\{\max(x)=x_d\}} \log\left(\frac{\delta'(x_d)}{d} h^{-1+1/d}(x_d) \right) dx,$$
$$J_3(\varepsilon) = d \int_{[\varepsilon,1-\varepsilon]^d} c_\delta(x) \mathbf{1}_{\{\max(x)=x_d\}} \left((d-1)F(x_d) - \sum_{i=1}^{d-1} F(x_i) \right) dx.$$

We introduce $A_\varepsilon(r) = \int_\varepsilon^r a(x)\,dx$. For $J_1(\varepsilon)$, we have:

$$\begin{aligned} J_1(\varepsilon) &= d(d-1) \int_{[\varepsilon,1-\varepsilon]^d} \mathbf{1}_{\{\max(x)=x_d\}} b(x_d) \prod_{j=1}^{d-1} a(x_j) \log\left(w(x_1)\right)\, dx \\ &= d(d-1) \int_{[\varepsilon,1-\varepsilon]} \left(\int_{[t,1-\varepsilon]} A_\varepsilon^{d-2}(s) b(s)\, ds \right) a(t) \log\left(w(t)\right)\, dt. \end{aligned}$$

Notice that using (2.11) and (2.13), we have:

$$\begin{aligned} \int_{[t,1-\varepsilon]} A_\varepsilon^{d-2}(s) b(s)\, ds &= \int_{[t,1]} A^{d-2}(s) b(s)\, ds - \int_{[t,1]} \left(A^{d-2}(s) - A_\varepsilon^{d-2}(s) \right) b(s)\, ds \\ &\quad - \int_{[1-\varepsilon,1]} A_\varepsilon^{d-2}(s) b(s)\, ds. \\ &= \frac{h(t)}{(d-1)A(t)} - \int_t^1 \left(A^{d-2}(s) - A_\varepsilon^{d-2}(s) \right) b(s)\, ds \\ &\quad - \int_{[1-\varepsilon,1]} A_\varepsilon^{d-2}(s) b(s)\, ds. \end{aligned}$$

By Fubini's theorem, we get:

$$J_1(\varepsilon) = J_{1,1}(\varepsilon) - J_{1,2}(\varepsilon) - J_{1,3}(\varepsilon),$$

with:

$$J_{1,1}(\varepsilon) = \int_{[\varepsilon,1-\varepsilon]} (d - \delta'(t)) \log(w(t))\, dt$$
$$J_{1,2}(\varepsilon) = d(d-1) \left(\int_{[1-\varepsilon,1]} A_\varepsilon^{d-2}(s) b(s)\, ds \right) \int_{[\varepsilon,1-\varepsilon]} a(t) \log(w(t))\, dt$$
$$J_{1,3}(\varepsilon) = d(d-1) \int_{[\varepsilon,1-\varepsilon]} \left(\int_t^1 \left(A^{d-2}(s) - A_\varepsilon^{d-2}(s) \right) b(s)\, ds \right) a(t) \log(w(t))\, dt.$$

To study $J_{1,2}$, we first give an upper bound for the term $\int_{[1-\varepsilon,1]} A_\varepsilon^{d-2}(s) a(s) b(s)\, ds$:

$$\begin{aligned} \int_{[1-\varepsilon,1]} A_\varepsilon^{d-2}(s) b(s)\, ds &\le \int_{[1-\varepsilon,1]} A^{d-2}(s) b(s)\, ds \\ &= \frac{1}{(d-1)}\, h^{1-1/d} (1-\varepsilon)\, \mathrm{e}^{-F(1-\varepsilon)} \\ &\le (d-1)^{-1/d} \varepsilon^{1-1/d}, \end{aligned} \tag{2.39}$$

where we used that $A_\varepsilon(s) \le A(s)$ for $s > \varepsilon$ for the first inequality, (2.13) for the first equality, and (2.36) for the last inequality. Since $t \log(t) \ge -1/\mathrm{e}$, we have, using (2.37):

$$\begin{aligned} J_{1,2}(\varepsilon) &\ge -\frac{d(d-1)}{\mathrm{e}} \left(\int_{[1-\varepsilon,1]} A_\varepsilon^{d-2}(s) b(s)\, ds \right) \int_{[\varepsilon,1-\varepsilon]} \mathrm{e}^{F(t)}\, dt \\ &\ge -\frac{d}{\mathrm{e}} h^{1-1/d} (1-\varepsilon) \int_{[\varepsilon,1-\varepsilon]} \mathrm{e}^{F(t)-F(1-\varepsilon)}\, dt \\ &\ge -\frac{d}{\mathrm{e}} ((d-1)\varepsilon)^{1-1/d}, \end{aligned}$$

where we used (2.13) for the second inequality, and that F is non-decreasing and (2.39) for the third inequality. On the other hand, we have $t \log(t) \le t^{\frac{1}{1-1/d}}$, if $t \ge 0$, which gives:

$$\begin{aligned} J_{1,2}(\varepsilon) &\le d(d-1) \left(\int_{[1-\varepsilon,1]} A_\varepsilon^{d-2}(s) b(s)\, ds \right) \int_{[\varepsilon,1-\varepsilon]} \mathrm{e}^{F(t)} \frac{\left(\frac{d-\delta'(t)}{d}\right)^{\frac{1}{1-1/d}}}{h(t)}\, dt \\ &= d h^{1-1/d} (1-\varepsilon) \int_{[\varepsilon,1-\varepsilon]} \frac{\mathrm{e}^{F(t)-F(1-\varepsilon)}}{h(t)}\, dt \\ &= d h^{1-1/d} (1-\varepsilon) \left(1 - \mathrm{e}^{F(\varepsilon)-F(1-\varepsilon)} \right) \\ &\le d((d-1)\varepsilon)^{1-1/d}, \end{aligned}$$

where we used (2.39) and $t^{\frac{1}{1-1/d}} \le 1$ for $t \in I$ for the first inequality, and that F is non-decreasing for the last. This proves that $\lim_{\varepsilon \to 0} J_{1,2}(\varepsilon) = 0$. For $J_{1,3}(\varepsilon)$, we first observe that for $s \in [\varepsilon, 1-\varepsilon]$ we have $A_\varepsilon(s) \le A(s)$ and thus:

$$\left(A^{d-2}(s) - A_\varepsilon^{d-2}(s) \right) = A(\varepsilon) \sum_{i=0}^{d-3} A^i(s) A_\varepsilon^{d-3-i}(s) \le (d-2) A(\varepsilon) A^{d-3}(s). \tag{2.40}$$

Using the previous inequality we obtain:

$$\begin{aligned}
J_{1,3}(\varepsilon) &= d(d-1)\int_{[\varepsilon,1-\varepsilon]}\left(\int_t^1\left(A^{d-2}(s)-A_\varepsilon^{d-2}(s)\right)b(s)\,ds\right)a(t)\log\left(w(t)\right)\,dt\\
&\geq -\frac{d(d-1)}{\mathrm{e}}\int_{[\varepsilon,1-\varepsilon]}\left(\int_t^1\left(A^{d-2}(s)-A_\varepsilon^{d-2}(s)\right)b(s)\,ds\right)\mathrm{e}^{F(t)}\,dt\\
&\geq -\frac{d(d-1)(d-2)A(\varepsilon)}{\mathrm{e}}\int_{[\varepsilon,1-\varepsilon]}\left(\int_t^1 A^{d-3}(s)b(s)\,ds\right)\mathrm{e}^{F(t)}\,dt\\
&\geq -\frac{d(d-1)(d-2)A(\varepsilon)}{\mathrm{e}}\int_{[\varepsilon,1-\varepsilon]}\frac{\left(\int_t^1 A^{d-2}(s)b(s)\,ds\right)}{A(t)}\mathrm{e}^{F(t)}\,dt\\
&= -\frac{d(d-2)A(\varepsilon)}{\mathrm{e}}\int_{[\varepsilon,1-\varepsilon]}\frac{h(t)}{A^2(t)}\mathrm{e}^{F(t)}\,dt\\
&= -\frac{d(d-2)h^{1/d}(\varepsilon)}{\mathrm{e}}\int_{[\varepsilon,1-\varepsilon]}h(t)^{1-2/d}\,\mathrm{e}^{F(\varepsilon)-F(t)}\,dt\\
&\geq -\frac{d(d-2)(d-1)^{1-1/d}\varepsilon^{1/d}}{\mathrm{e}},
\end{aligned}$$

where we used $t\log(t)\geq -1/\mathrm{e}$ for the first inequality, (2.40) for the second, (2.11) and (2.13) in the following equality, and (2.36) to conclude. For an upper bound, we have after noticing that $t\log(t)\leq t^2$:

$$\begin{aligned}
J_{1,3}(\varepsilon) &= d(d-1)\int_{[\varepsilon,1-\varepsilon]}\left(\int_t^1\left(A^{d-2}(s)-A_\varepsilon^{d-2}(s)\right)b(s)\,ds\right)a(t)\log\left(w(t)\right)\,dt\\
&\leq d(d-1)\int_{[\varepsilon,1-\varepsilon]}\left(\int_t^1\left(A^{d-2}(s)-A_\varepsilon^{d-2}(s)\right)b(s)\,ds\right)\mathrm{e}^{F(t)}\,w^2(t)\,dt\\
&\leq d(d-1)(d-2)A(\varepsilon)\int_{[\varepsilon,1-\varepsilon]}\frac{\left(\int_t^1 A^{d-2}(s)b(s)\,ds\right)}{A(t)}\mathrm{e}^{F(t)}\,h^{-2+2/d}(t)\,dt\\
&= d(d-2)A(\varepsilon)\int_{[\varepsilon,1-\varepsilon]}\frac{\mathrm{e}^{-F(t)}}{h(t)}\,dt\\
&= d(d-2)h^{1/d}(\varepsilon)(1-\mathrm{e}^{F(\varepsilon)-F(1-\varepsilon)})\\
&\leq d(d-2)(d-1)^{1/d}\varepsilon^{1/d},
\end{aligned}$$

where we used (2.40) and $0\leq(d-\delta'(t))/d\leq 1$ for the second inequality; (2.11) and (2.13) in the second equality; and (2.36) to conclude. The results on the two bounds show that $\lim_{\varepsilon\to 0}J_{1,3}(\varepsilon)=0$. Similarly, for $J_2(\varepsilon)$, we get:

$$\begin{aligned}
J_2(\varepsilon) &= \int_{[\varepsilon,1-\varepsilon]^d}\mathbf{1}_{\{\max(x)=x_d\}}b(x_d)\prod_{j=1}^{d-1}a(x_j)\log\left(\frac{\delta'(x_d)}{d}h^{-1+1/d}(x_d)\right)\,dx\\
&= d\int_{[\varepsilon,1-\varepsilon]}A_\varepsilon^{d-1}(t)b(t)\log\left(\frac{\delta'(t)}{d}h^{-1+1/d}(t)\right)\,dt\\
&= d\int_{[\varepsilon,1-\varepsilon]}A^{d-1}(t)b(t)\log\left(\frac{\delta'(t)}{d}h^{-1+1/d}(t)\right)\,dt\\
&\qquad - d\int_{[\varepsilon,1-\varepsilon]}\left(A^{d-1}(t)-A_\varepsilon^{d-1}(t)\right)b(t)\log\left(\frac{\delta'(t)}{d}h^{-1+1/d}(t)\right)\,dt\\
&= J_{2,1}(\varepsilon)-J_{2,2}(\varepsilon)
\end{aligned}$$

with $J_{2,1}(\varepsilon)$ and $J_{2,2}(\varepsilon)$ given by, using (2.12):

$$J_{2,1}(\varepsilon) = d\int_{[\varepsilon,1-\varepsilon]} A^{d-1}(t)b(t)\log\left(\frac{\delta'(t)}{d}h^{-1+1/d}(t)\right)\,dt$$
$$J_{2,2}(\varepsilon) = d\int_{[\varepsilon,1-\varepsilon]} \left(A^{d-1}(t) - A_\varepsilon^{d-1}(t)\right)b(t)\log\left(\frac{\delta'(t)}{d}h^{-1+1/d}(t)\right)\,dt.$$

By (2.12), we have:

$$J_{2,1}(\varepsilon) = \int_{[\varepsilon,1-\varepsilon]} \delta'(t)\log\left(\frac{\delta'(t)}{d}h^{-1+1/d}(t)\right)\,dt. \tag{2.41}$$

Similarly to $J_{1,3}(\varepsilon)$ we can show that $\lim_{\varepsilon\to 0} J_{2,2}(\varepsilon) = 0$.

Adding up $J_1(\varepsilon)$ and $J_2(\varepsilon)$ gives

$$J_1(\varepsilon) + J_2(\varepsilon) = \mathcal{J}_\varepsilon(\delta) + J_4(\varepsilon) - d\log(d)(1-2\varepsilon) - J_{1,2}(\varepsilon) - J_{1,3}(\varepsilon) - J_{2,2}(\varepsilon)$$

with

$$\mathcal{J}_\varepsilon(\delta) = (d-1)\int_\varepsilon^{1-\varepsilon} |\log(h(t))|\,dt,$$
$$J_4(\varepsilon) = \int_\varepsilon^{1-\varepsilon} (d-\delta'(t))\log(d-\delta'(t))\,dt + \int_\varepsilon^{1-\varepsilon} \delta'(t)\log(\delta'(t))\,dt.$$

Notice that $\mathcal{J}_\varepsilon(\delta)$ is non-decreasing in $\varepsilon > 0$ and that:

$$\mathcal{J}(\delta) = \lim_{\varepsilon\to 0} \mathcal{J}_\varepsilon(\delta).$$

Since $\delta'(t) \in [0,d]$, we deduce that $(d-\delta')\log(d-\delta')$ and $\delta'\log(\delta')$ are bounded on I from above by $d\log(d)$ and from below by $-1/\mathrm{e}$ and therefore integrable on I. This implies :

$$\lim_{\varepsilon\to 0} J_4(\varepsilon) = \mathcal{I}_1(\delta') + \mathcal{I}_1(d-\delta').$$

As for $J_3(\varepsilon)$, we have by integration by parts:

$$\begin{aligned}
J_3(\varepsilon) &= d\int_{[\varepsilon,1-\varepsilon]^d} \mathbf{1}_{\{\max(x)=x_d\}} b(x_d)\prod_{i=i}^{d-1} a(x_i)\left((d-1)F(x_d) - \sum_{i=1}^{d-1} F(x_i)\right)dx \\
&= d(d-1)\int_{[\varepsilon,1-\varepsilon]} A_\varepsilon^{d-1}(t)b(t)F(t)\,dt \\
&\qquad - d(d-1)\int_{[\varepsilon,1-\varepsilon]} A_\varepsilon^{d-2}(t)b(t)\left(\int_\varepsilon^t a(s)F(s)\,ds\right)dt \\
&= d(d-1)\int_{[\varepsilon,1-\varepsilon]} A_\varepsilon^{d-1}(t)b(t)F(t)\,dt \\
&\qquad - d(d-1)\int_{[\varepsilon,1-\varepsilon]} A_\varepsilon^{d-2}(t)b(t)\left(A_\varepsilon(t)F(t) - \frac{d-1}{d}\int_\varepsilon^t \frac{A_\varepsilon(s)}{h(s)}\,ds\right)dt \\
&= (d-1)^2\int_{[\varepsilon,1-\varepsilon]} \left(\int_t^{1-\varepsilon} A_\varepsilon^{d-2}(s)b(s)\,ds\right)\frac{A_\varepsilon(t)}{h(t)}\,dt.
\end{aligned}$$

By the monotone convergence theorem, (2.11) and (2.13) we have:

$$\lim_{\varepsilon\to 0} J_3(\varepsilon) = (d-1)^2\int_I \left(\int_t^1 A^{d-2}(s)b(s)\,ds\right)\frac{A(t)}{h(t)}\,dt = d-1.$$

Summing up all the terms and taking the limit $\varepsilon = 0$ give :

$$\begin{aligned}
\mathcal{I}(C_\delta) &= (d-1)\int_I |\log(t-\delta(t))|\,dt + \mathcal{I}_1(\delta') + \mathcal{I}_1(d-\delta') - d\log(d) - (d-1) \\
&= (d-1)\mathcal{J}(\delta) + \mathcal{G}(\delta).
\end{aligned}$$

2.8.2 Proof of Lemma 2.14

Set $\bar{\Phi}(x) = 1 - \Phi(x)$, the survival function of the standard Gaussian distribution. We recall the well known approximation of $\bar{\Phi}(t)$ for $t > 0$:

$$\bar{\Phi}(t) \leq \frac{\varphi(t)}{t} \quad \text{and} \quad \bar{\Phi}(t) = \frac{\varphi(t)}{t}\left(1 - \frac{1}{t^2} + g(t)\right) \quad \text{with} \quad 0 \leq g(t) \leq \frac{3}{t^4}. \tag{2.42}$$

We set $W = (X_1 - \rho X_2)/\sqrt{1-\rho^2}$ so that W is standard normal and independent of X_2. We have:

$$\mathbb{P}(X_1 \geq \alpha t | X_2 = t) = \mathbb{P}\left(W \geq \frac{(\alpha-\rho)t}{\sqrt{1-\rho^2}}\right) = \bar{\Phi}\left(\frac{(\alpha-\rho)t}{\sqrt{1-\rho^2}}\right)$$

Since $\alpha \geq \rho$, this gives:

$$\mathbb{P}(X_1 \geq \alpha t | X_2 = t) = \frac{1}{\sqrt{2\pi}\, t} \frac{\sqrt{1-\rho\hat{\mathrm{A}}^2}}{\alpha - \rho}(1 + O(t^{-2})) \exp\left(-\frac{1}{2}\frac{(\alpha-\rho)^2}{(1-\rho\hat{\mathrm{A}}^2)}t^2\right). \tag{2.43}$$

For (Y_1, Y_2), we have using notation from Section 2.2:

$$\mathbb{P}(Y_1 \geq \alpha t | Y_2 = t) = \int_{\alpha t}^{\infty} c_\delta(\Phi(x), \Phi(t))\varphi(x)\, dx = \int_{\Phi(\alpha t)}^{1} b(s)a(\Phi(t))\, ds = B(\Phi(\alpha t))a(\Phi(t)),$$

with B defined for $r \in I$ as $B(r) = \int_r^1 b(s)\, ds$. Using that $B(r) = h^{1/2}(r)\, \mathrm{e}^{-F(r)}$ as well as the formulas (2.33) for δ_ρ and δ'_ρ, elementary computations give:

$$\mathbb{P}(Y_1 \geq \alpha t | Y_2 = t) = \bar{\Phi}\left(\sqrt{\frac{1-\rho}{1+\rho}}t\right) \mathrm{e}^{-\Gamma_t}, \tag{2.44}$$

with

$$\Gamma_t = \int_t^{\alpha t} \Phi\left(\sqrt{\frac{1-\rho}{1+\rho}}u\right) \frac{\varphi(u)}{\bar{\Phi}(u) - \bar{\Phi}_\rho(u,u)}\, du \quad \text{and} \quad \bar{\Phi}_\rho(u,v) = \mathbb{P}(X_1 \geq u, X_2 \geq v).$$

Using (2.42), it is easy to check that $\bar{\Phi}_\rho(u,u) = O(\varphi(u)u^{-5})$ for u large, and deduce that:

$$\Gamma_t = \frac{(\alpha^2-1)t^2}{2} + \log(\alpha) + O\left(t^{-2}\right).$$

Using (2.44) and (2.42), we get:

$$\mathbb{P}(Y_1 \geq \alpha t | Y_2 = t) = \frac{1}{\sqrt{2\pi}\, t}\, \frac{1}{\alpha}\sqrt{\frac{1+\rho}{1-\rho}}\,(1 + O(t^{-2})) \exp\left(-\frac{1}{2}\left(\frac{1-\rho}{1+\rho} + \alpha^2 - 1\right)t^2\right).$$

Using (2.43), we obtain (2.34).

2.9 Supplementary material

We give the proof of Proposition 2.4 part (a), see Lemma 2.20.

Let $T_i = \{u \in I^d; \max(u) = u_i\}$ for $1 \leq i \leq d$. For $x \leq y$ elements of $\mathbb{R}^d$, we consider the hyper-rectangle $[x, y] = \{z \in \mathbb{R}^d, x \leq z \leq y\}$. The next Lemma ensures that every symmetric feasible solution of (P^δ) which is not of the form described in Lemma 2.11 can be changed locally on any hyper-rectangle subset of T_1 (and by symmetry on all T_i), in order to conserve or increase its Kullback-Leibler divergence.

Lemma 2.16. *Let $C \in \mathcal{C}^\delta$ with density c be a symmetric feasible solution to (P^δ). Let $x = (x_1, \dots, x_d)$ and $y = (y_1, \dots, y_d)$ be elements of T_1 such that $x \le y$, $x_i = x_2$ and $y_i = y_2$ for all $2 \le i \le d$. Then we can define non-negative measurable functions $\tilde{a}, \tilde{b}$ such that $\tilde{c}$ defined by $\tilde{c} = c$ on $I^d \setminus [x, y]$ and a.e. for $u = (u_1, \dots, u_d) \in [x, y]$:*

$$\tilde{c}(u) = \tilde{b}(u_1) \prod_{i=2}^{d} \tilde{a}(u_i), \tag{2.45}$$

is the density of a copula $\tilde{C}$ which verifies $\tilde{C} \in \mathcal{C}^\delta$ and $\mathcal{I}(\tilde{C}) \le \mathcal{I}(C)$.

Proof. For $u = (u_1, \dots, u_d) \in I^d$, we set $u_{(-i)} = (u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_d) \in I^{d-1}$. Let $M = (\int_{[x,y]} c(u)\, du)^{1/d}$. If $M = 0$, then simply take $\tilde{a} = 0$ and $\tilde{b} = 0$ and the proof is complete. If $M > 0$, we define the functions $\tilde{b}$ and $\tilde{a}_i$, $2 \le i \le d$ as:

$$\tilde{b}(u_1) = M^{1-d} \int_{[x_2,y_2]^{d-1}} c(u)\, du_{(-1)} \quad \text{for } u_1 \in [x_1, y_1], \tag{2.46}$$

$$\tilde{a}_i(u_i) = M^{1-d} \int_{[x_1,y_1]\times[x_2,y_2]^{d-2}} c(u)\, du_{(-i)} \quad \text{for } u_i \in [x_2, y_2]. \tag{2.47}$$

Notice that these functions are non-negative, and we have:

$$\int_{x_1}^{y_1} \tilde{b}(u_1)\, du_1 = M, \quad \text{and} \quad \int_{x_2}^{y_2} \tilde{a}_i(u_i)\, du_i = M \quad \text{for } 2 \le i \le d. \tag{2.48}$$

By the symmetry of c and the integration domain for $2 \le i \le d$, we can deduce that $\tilde{a}_i = \tilde{a}$ for all $2 \le i \le d$. Let $\tilde{c}$ be defined by (2.45). We first check that $\mathcal{A}_{d+1}(\tilde{c}) = b_{d+1}$. Notice that $\mathcal{A}_{d+1}(\tilde{c})(r) = b_{d+1}(r)$ holds for $r \in [0, x_1]$, since the density has not been changed in the region of integration. When $r \in [x_1, 1]$, we have:

$$\mathcal{A}_{d+1}(\tilde{c})(r) = \int_{I^d} \tilde{c}(u) \mathbf{1}_{\{\max(u) \le r\}}\, du = \mathcal{A}_{d+1}(c)(r) + \int_{[x,y]} (\tilde{c}(u) - c(u)) \mathbf{1}_{\{\max(u) \le r\}}\, du. \tag{2.49}$$

Since we supposed that $[x, y] \subset T_1$, we have that $\max(u) = u_1$ for $u \in [x, y]$, in particular $y_2 \in [0, x_1]$ and thus $[x_2, y_2] \subset [0, x_1] \subset [0, r]$. Therefore, we get:

$$\begin{aligned}
\int_{[x,y]} \tilde{c}(u) \mathbf{1}_{\{\max(u) \le r\}}\, du &= \left(\prod_{i=2}^{d} \int_{x_2}^{y_2} \tilde{a}(u_i)\, du_i \right) \left(\int_{x_1}^{r} \tilde{b}(u_1)\, du_1 \right) \\
&= M^{d-1} \left(\int_{x_1}^{r} \tilde{b}(u_1)\, du_1 \right) \\
&= \int_{[x,y]} c(u) \mathbf{1}_{\{u_1 \le r\}}\, du, \\
&= \int_{[x,y]} c(u) \mathbf{1}_{\{\max(u) \le r\}}\, du,
\end{aligned}$$

where we used (2.46) for the third equality and $[x, y] \subset T_1$ for the last. This implies the last integral in (2.49) equals 0, ensuring $\mathcal{A}_{d+1}(\tilde{c})(r) = b_{d+1}(r)$ for all $r \in I$. Similarly, it is easy to check $\mathcal{A}_i(\tilde{c}) = b_i$ for $1 \le i \le d$.

To show that $\mathcal{I}(\tilde{C}) \le \mathcal{I}(C)$ that is $\mathcal{I}_d(\tilde{c}) \le \mathcal{I}_d(c)$, we consider the following optimization problem, say $(P^{[x,y]})$, minimize $\mathcal{I}_{[x,y]}(f)$ subject to:

$$\begin{cases}
\int_{[x,y]} f(u) \mathbf{1}_{\{u_1 \le r\}}\, du = M^{-d} \int_{x_1}^{r} \tilde{b}(s)\, ds, & \text{for all } r \in [x_1, y_1], \\
\int_{[x,y]} f(u) \mathbf{1}_{\{u_i \le r\}}\, du = M^{-d} \int_{x_2}^{r} \tilde{a}(s)\, ds, & \text{for all } 2 \le i \le d, r \in [x_2, y_2], \\
f \ge 0 \text{ a.e. and } f \in L^1([x, y]),
\end{cases} \tag{2.50}$$

where $\mathcal{I}_{[x,y]}(f) = \int_{[x,y]} f \log(f)$. This problem is exactly the problem of finding the density f of the maximum entropy distribution on $[x, y]$ with fixed marginals given by their densities $M^{-1}\tilde{b}(u_1)\mathbf{1}_{[x_1,y_1]}(u_1)$ and $M^{-1}\tilde{a}(u_i)\mathbf{1}_{[x_2,y_2]}(u_i)$ for $2 \le i \le d$. These marginals verify

$\mathcal{I}_1(M^{-1}\tilde{b}\mathbf{1}_{[x_1,y_1]}) < +\infty$ and $\mathcal{I}_1(M^{-1}\tilde{a}\mathbf{1}_{[x_2,y_2]}) < +\infty$, since $\mathcal{I}_{[x,y]}(M^{-1}c) \leq \int_{I^d} c|\log(c)| < \infty$. Therefore by Corollary 3.2 of [48], $M^{-d}\tilde{c}$ is the optimal solution for $(P^{[x,y]})$, and in particular, this yields

$$\int_{[x,y]} \tilde{c}\log(\tilde{c}) \leq \int_{[x,y]} c\log(c),$$

ensuring $\mathcal{I}_d(\tilde{c}) \leq \mathcal{I}_d(c)$.

□

An important consequence of Lemma 2.16 and Proposition 2.7 is that if there exists an optimal (symmetric) solution of (P^δ), it has a product form on all hyper-rectangles included in T_1.

Corollary 2.17. *Assume there exists a feasible solution to (P^δ). Let $x = (x_1, \dots, x_d)$ and $y = (y_1, \dots, y_d)$ be elements of T_1 such that $x \leq y$, $x_i = x_2$ and $y_i = y_2$ for all $2 \leq i \leq d$. Then, there exist non-negative measurable functions g, h such that the unique optimal solution c^* of (P^δ) takes the form $c^*(u) = g(u_1)\prod_{i=2}^d h(u_i)$ for a.e. $u = (u_1, \dots, u_d) \in [x, y]$.*

For arbitrary sets $A, B \subset \mathbb{R}^d$, we write $|A|$ the Lebesgue measure of A, $A \subset B$ a.e. if $|A \bigcap B^c| = 0$, $A = B$ a.e. if a.e. $A \subset B$ and a.e. $B \subset A$. For A, B subsets of I such, we define:

$$\Theta(A, B) = T_1 \cap \left(\left(A \times I^{d-1} \right) \cup \left(\bigcup_{i=2}^{d} I^{i-1} \times B \times I^{d-i} \right) \right). \tag{2.51}$$

We say that a subset of T_1 is a stripe if it is a.e. equal to $\Theta(A, B)$ for some A, B Borel subsets of I.

We give a characterization of zeros of any optimal solution of (P^δ).

Lemma 2.18. *Assume there exists an optimal solution c^* of (P^δ). Let $K = \{u \in T_1, c^*(u) = 0\}$ denote the subset of T_1 where c^* vanishes. Then at least one of the two following statements hold:*

- *The set K is a stripe.*
- *There exists $t \in (0, 1)$ such that a.e. $(T_1 \bigcap ([t, 1] \times I^{d-1}) \setminus [t, 1]^d) \subset K$.*

Proof. For $t \in (0, 1)$, the hyper-rectangle $R_t = [t, 1] \times [0, t]^{d-1}$ is a subset of T_1. We use notation of Corollary 2.17 with $x_1 = t$, $x_2 = 0$, $y_1 = 1$, $y_2 = t$, and set $A_t = \{s \in [t, 1],\ g(s) = 0\}$ as well as $B_t = \{s \in [0, t],\ h(s) = 0\}$. We consider the stripe $\Theta_t = \Theta(A_t, B_t)$. Notice that by construction a.e.:

$$K \bigcap R_t = \Theta_t \bigcap R_t.$$

We shall distinguishes two cases. In the first case, we assume that for all $q \in \mathbb{Q} \bigcap (0, 1)$, a.e. $\Theta_q \subset K$. Set $A = \bigcup_{q \in \mathbb{Q} \bigcap (0,1)} A_q$, $B = \bigcup_{q \in \mathbb{Q} \bigcap (0,1)} B_q$, $\Theta = \Theta(A, B)$ so that $\Theta = \bigcup_{q \in \mathbb{Q} \bigcap (0,1)} \Theta_q \subset K$. We get a.e.:

$$\Theta \bigcap R_q \subset K \bigcap R_q = \Theta_q \bigcap R_q \subset \Theta \bigcap R_q.$$

That is a.e. for all $q \in \mathbb{Q} \bigcap (0, 1)$, $K \bigcap R_q = \Theta \bigcap R_q$. Use that a.e. $T_1 = \bigcup_{q \in \mathbb{Q} \bigcap (0,1)} R_q$ to get that a.e. $\Theta = K$.

In the second case, we assume there exists $q \in \mathbb{Q} \bigcap (0, 1)$ such that $|\Theta_q \bigcap K^c| > 0$. We first assume that:

$$|(A_q \times I^{d-1}) \bigcap K^c| > 0.$$

We define:

$$t = \inf\{s > q,\ |(A_q \times I^{d-1}) \bigcap K^c \bigcap R_s| > 0\}.$$

Notice that t belongs to $[q, 1)$ as $T_1 = \bigcup_{s \in \mathbb{Q} \bigcap (0,1)} R_s$ and the boundary of T_1 has zero Lebesgue measure. By continuity, we get for all $\varepsilon > 0$ small enough:

$$|(A_q \times I^{d-1}) \bigcap K^c \bigcap R_t| = 0 \quad \text{and} \quad |(A_q \times I^{d-1}) \bigcap K^c \bigcap R_{t+\varepsilon}| > 0.$$

We deduce from the last equality that $A_q \bigcap [t,1] \subset A_t$. We deduce from the last inequality and the representation of Corollary 2.17 for the hyper-rectangle $R_{t+\varepsilon}$ that g is non-zero on a subset of $A_q \bigcap [t+\varepsilon, 1]$ of positive Lebesgue measure. Since $A_q \bigcap [t,1] \subset A_t$ and c^* is zero on $A_t \times [0,t]^{d-1}$ (by definition of A_t), this implies that $h = 0$ on $[0,t]$, that is a.e. $[t+\varepsilon, 1] \times [0,t]^{d-1} \subset K$. Let ε goes down to 0, we deduce that a.e. $R_t \subset K$.

Let $s > t$. Using the representation of Corollary 2.17 on the hyper-rectangle R_s, since $c^* = 0$ on R_t, we get either $g = 0$ on $[s,1]$ or $h = 0$ at least on $[0,t]$ and thus $c^* = 0$ a.e. on $[s,1] \times [0,t] \times [0,s]^{d-2}$. By symmetry, and letting s run in $(t,1) \bigcap \mathbb{Q}$, we get a.e. $(T_1 \bigcap ([t,1] \times I^{d-1}) \setminus [t,1]^d) \subset K$.

If $|(A_q \times I^{d-1}) \bigcap K^c| = 0$, then we have $\sum_{i=2}^d |(I^{i-1} \times B_q \times I^{d-i}) \bigcap K^c| > 0$. This case can be handled similarly to the previous one. □

Lastly we show that if the optimal solution of (P^δ) vanishes on a stripe of T_1, then the stripe is a.e. a subset of Z_δ.

Lemma 2.19. *Assume there exists an optimal solution c^* of (P^δ). Let $K = \{u \in T_1, c^*(u) = 0\}$ denote the subset of T_1 where c^* vanishes. Let Θ be a stripe such that a.e. $\Theta \subset K$. Then we have a.e. $\Theta \subset Z_\delta$.*

Proof. Recall that a.e. $\Theta = \Theta(A,B)$ is defined by (2.51) with A and B Borel sets. Using that a.e. $\Theta \subset K$ and the symmetry of c^*, we get:

$$\int_{I^d} c^*(u) \mathbf{1}_A(\max(u))\, du = 0.$$

By the monotone class theorem and the constraint $\mathcal{A}_{d+1}(c^*) = \delta$, we obtain that:

$$0 = \int_{I^d} c^*(u) \mathbf{1}_A(\max(u))\, du = \int_A \delta'(s)\, ds,$$

that is $\delta' = 0$ a.e. on A, since δ' is non-negative. On the other hand we have:

$$\int_{T_1} c^*(u) \left(\sum_{i=2}^d \mathbf{1}_B(u_i) \right) du = 0.$$

By the symmetry of c^*, we get:

$$\begin{aligned} 0 &= \sum_{j=1}^d \left(\int_{T_j} c^*(u) \left(\sum_{i=1, i\neq j}^d \mathbf{1}_B(u_i) \right) du \right) \\ &= \sum_{i=1}^d \left(\int_{I^d} c^*(u) \mathbf{1}_B(u_i)\, du - \int_{I^d} c^*(u) \mathbf{1}_{\{u_i = \max(u)\}} \mathbf{1}_B(u_i)\, du \right) \\ &= \sum_{i=1}^d \left(\int_{I^d} c^*(u) \mathbf{1}_B(u_i)\, du \right) - \int_{I^d} c^*(u) \mathbf{1}_B(\max(u))\, du. \end{aligned}$$

Applying the monotone class theorem and the constraints $\mathcal{A}_i(c^*) = b_i$, $1 \leq i \leq d+1$, we obtain that:

$$0 = \sum_{i=1}^d \left(\int_{I^d} c^*(u) \mathbf{1}_B(u_i)\, du \right) - \int_{I^d} c^*(u) \mathbf{1}_B(\max(u))\, du = \int_B (d - \delta'(s))\, ds,$$

that is $\delta' = d$ a.e. on B since a.e. $\delta' \leq d$. This and $\delta' = 0$ a.e. on A implies that a.e. $\Theta \subset Z^\delta$. □

The next Lemma corresponds to Proposition 2.12 part (a).

Lemma 2.20. *Let $\delta \in \mathcal{D}_0$ such that $\Sigma_\delta = \{0,1\}$. If $\mathcal{J}(\delta) = +\infty$ then there exists no feasible solution to (P^δ).*

Proof. Let us assume that there exists a feasible solution to (P^δ). Then by Proposition 2.7, there exists a unique symmetric copula C^* with density c^* such that $\mathcal{I}(C^*) = \min_{C\in\mathcal{C}^\delta} \mathcal{I}(C) < +\infty$. Let $\Upsilon = \{u \in I^d, c^*(u) = 0\}$. By Proposition 2.8, we have that a.e. $Z_\delta \subset \Upsilon$.

By Proposition 2.2, $\mathcal{J}(\delta) = +\infty$ implies that $\mathcal{I}(C_\delta) = +\infty$, therefore c_δ is not a feasible solution to (P^δ). We deduce from Proposition 2.12 that $|\Upsilon \bigcap Z_\delta^c| > 0$. Since c^* i symmetric, this implies that, with $K = \Upsilon \cap T_1$, we have $|K \bigcap Z_\delta^c| > 0$. According to Lemma 2.19 this implies that K is not a stripe. We deduce from Lemma 2.18, that there exists $t \in (0, 1)$ such that a.e. $(T_1 \bigcap([t, 1] \times I^{d-1}) \setminus [t, 1]^d) \subset K$. By symmetry, we deduce that $c^* = 0$ on $I^d \setminus ([0, t]^d \bigcup [t, 1]^d)$. This in turn implies that $t \in \Sigma_\delta$. This leads to a contradiction since, we assumed that $\Sigma_\delta = \{0, 1\}$.

In conclusion we get there is no feasible solution to (P^δ). □

Chapter 3

Maximum entropy distribution of order statistics with given marginals

3.1 Introduction

Order statistics, an almost surely non-decreasing sequence of random variables, have received a lot of attention due to the diversity of possible applications. If $X = (X_1, \ldots, X_d)$ is a d-dimensional random vector, then its order statistics $X^{OS} = (X_{(1)}, \ldots, X_{(d)})$ corresponds to the permutation of the components of X in the non-decreasing order, so that $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(d)}$. The components of the underlying random vector X are usually, but not necessarily, independent and identically distributed (i.i.d.). Special attention has been given to extreme values $X_{(1)}$ and $X_{(d)}$, the range $X_{(d)} - X_{(1)}$, or the median value. Direct application of the distribution of the k-th largest order statistic occurs in various fields, such as climatology, extreme events, reliability, insurance, financial mathematics. We refer to the monographs of David and Nagaraja [55] and Arnold, Balakrishnan, and Nagaraja [5] for a general overview on the subject of order statistics. We are interested in the dependence structure of order statistics, which has received great attention. In the i.i.d. case, Bickel [21] showed that any two order statistics are positively correlated. The copula of the joint distribution of $X_{(1)}$ and $X_{(d)}$ is derived in Schmitz [156] with exact formulas for Kendall's τ and Spearman's ρ. In Avérous, Genest, and Kochar [9], it is shown that the dependence of the j-th order statistic on the i-th order statistic decreases as the distance between i and j increases according to the bivariate monotone regression dependence ordering. The copula connecting the limit distribution of the two largest order statistics, called bi-extremal copula, is given by de Melo Mendes and Sanfins [59] with some additional properties. Exact expressions for Pearson's correlation coefficient, Kendall's τ and Spearman's ρ for any two order statistics are obtained in Navarro and Balakrishnan [130]. For the non i.i.d. case, Kim and David [114] shows that some pairs of order statistics can be negatively correlated, if the underlying random vector is sufficiently negatively dependent. Positive dependence measures for two order statistics are considered in Boland, Hollander, Joag-Dev, and Kochar [24] when the underlying random variables are independent but arbitrarily distributed or when they are identically distributed but not independent. A generalization of these results for multivariate dependence properties is given by Hu and Chen [99]. See also Dubhashi and Häggström [66] for conditional distribution of order statistics.

Here, we focus on the cumulative distribution function (cdf) of order statistics without referring to an underlying distribution. That is, we consider random vectors $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ such that a.s. $X_1 \leq \cdots \leq X_d$ and we suppose that the one-dimensional marginal distributions $\mathbf{F} = (\mathbf{F}_i, 1 \leq i \leq d)$ are given, where $\mathbf{F}_i$ is the cdf of X_i. A necessary and sufficient condition for the existence of a joint distribution of order statistics with one-dimensional marginals $\mathbf{F}$ is that they are stochastically ordered, that is $\mathbf{F}_{i-1}(x) \geq \mathbf{F}_i(x)$ for all $2 \leq i \leq d, x \in \mathbb{R}$. With the marginals fixed, the joint distribution of the order statistics can be characterized by the connecting copula of the random vector, which contains all information on the dependence

structure of the order statistics. Copulas of order statistics derived from an underlying i.i.d. sample were considered in [9] in order to calculate measures of concordance between any two pairs of order statistics. For order statistics derived from a general parent distribution, Navarro and Spizzichino [131] shows that the copula of the order statistics depends on the marginals and the copula of the parent distribution through an exchangeable copula and the average of the marginals. Construction of some copula of order statistics with given marginals were given in Lebrun and Dutfoy [119].

Our aim is to find the cdf of order statistics of dimension d with fixed marginals which maximizes the relative entropy H_h defined by (3.13). In an information-theoretic interpretation, the maximum entropy distribution is the least informative among order statistics with given marginals. This problem appears in models where the one-dimensional marginals are well known (either from different experimentation or from physical models) but the dependence structure is unknown, see Butucea, Delmas, Dutfoy, and Fischer [34]. In [33], the same authors gave, when it exists, the maximum entropy distribution of $(X_1, \dots, X_d)$ such that X_i is uniformly distributed on $[0, 1]$ for $1 \leq i \leq d$ and the distribution of $X_{(d)} = \max_{1 \leq i \leq d} X_i$ is given, see Remark 3.33.

For a d-dimensional random variable $X = (X_1, \dots, X_d)$ with cdf F and copula C_F, the relative entropy of F can be decomposed into the sum of the relative entropy (with respect to a one-dimensional probability density h) of its one-dimensional marginals plus the entropy of C_F, see Lemma3.1. In our case, since the marginals $\mathbf{F} = (\mathbf{F}_i, 1 \leq i \leq d)$ are fixed, maximizing the entropy of the joint distribution F of an order statistics is equivalent to maximizing the entropy of its copula C_F under constraints, (see Section 3.2.4). Therefore we shall find the maximum entropy copula for order statistics with fixed marginal distributions.

The main result of this paper is given by Theorem 3.37. It states that there exists a unique maximum entropy cdf $F_{\mathbf{F}}$ given by (3.52) if and only if:

$$\sum_{i=1}^{d} H_h(\mathbf{F}_i) - \sum_{i=2}^{d} \int_{\mathbb{R}} \mathbf{F}_i(dt) \; |\log\left(\mathbf{F}_{i-1}(t) - \mathbf{F}_i(t)\right)| > -\infty.$$

In this case $F_{\mathbf{F}}$ is absolutely continuous with density $f_{\mathbf{F}}$ defined as, for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$:

$$f_{\mathbf{F}}(x) = \mathbf{f}_1(x_1) \prod_{i=2}^{d} \frac{\mathbf{f}_i(x_i)}{\mathbf{F}_{i-1}(x_i) - \mathbf{F}_i(x_i)} \exp\left(- \int_{x_{i-1}}^{x_i} \frac{\mathbf{f}_i(s)}{\mathbf{F}_{i-1}(s) - \mathbf{F}_i(s)} \, ds \right) \mathbf{1}_{L^{\mathbf{F}}}(x),$$

where $\mathbf{f}_i$ is the density function of $\mathbf{F}_i$ and $L^{\mathbf{F}} \subset \mathbb{R}^d$ is the set of ordered vectors $(x_1, \dots, x_d)$, that is $x_1 \leq \cdots \leq x_d$, such that $\mathbf{F}_{i-1}(t) > \mathbf{F}_i(t)$ for all $t \in (x_{i-1}, x_i)$ and $2 \leq i \leq d$. See Example 3.41 for an illustrative example.

The rest of the paper is organized as follows. In Section 3.2, we introduce the basic notations and give the definition of the objects used in later parts. Section 3.3 describes the connection between copulas of order statistics with fixed marginals, and symmetric copulas with fixed multidiagonals. The multidiagonal, given by Definition 3.8, is the generalization of the diagonal section for copulas, which received great attention in copula literature. We show that there exist a one-to-one map between these two sets of copulas, see Corollary 3.16. This bijection has good properties with respect to the entropy as explained in Proposition 3.24. In Section 3.4, we determine the maximum entropy copula with fixed multidiagonal, see Theorem 3.32. Since we obtain a symmetric copula as a result, this is also the maximum entropy symmetric copula with fixed multidiagonal. In Section 3.5, we use the one-to-one map between the two sets of copulas established in Section 3.3 to give the maximum entropy copula of order statistics with fixed marginals . We finally obtain the density of the maximum entropy distribution for order statistics with fixed marginals by composing the maximum entropy copula with the marginals, see Theorem 3.37. Section 3.6 contains the detailed proofs of Theorem 3.32 and other results from Section 3.4. Section 3.7 collects the main notations of the paper to facilitate reading.

3.2 Notations and definitions

3.2.1 Notations in $\mathbb{R}^d$ and generalized inverse

For a Borel set $A \subset \mathbb{R}^d$, we write $|A|$ for its Lebesgue measure. For $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ and $y = (y_1, \dots, y_d) \in \mathbb{R}^d$, we write $x \leq y$ if $x_i \leq y_i$ for all $1 \leq i \leq d$. We define $\min x = \min\{x_i,\ 1 \leq i \leq d\}$ and $\max x = \max\{x_i,\ 1 \leq i \leq d\}$ for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$. If J is a real-valued function defined on $\mathbb{R}$, we set $J(x) = (J(x_1), \dots, J(x_d))$. We shall consider the following subsets of $\mathbb{R}^d$:

$$S = \{(x_1, \dots, x_d) \in \mathbb{R}^d, x_1 \leq \dots \leq x_d\} \quad \text{and} \quad \triangle = S \cap I^d,$$

with $I = [0, 1]$. In what follows, usually x, y will belong to $\mathbb{R}^d$, and s, t to $\mathbb{R}$ or I. For a set $A \subset \mathbb{R}$, we note by $A^c = \mathbb{R} \setminus A$ its complementary set.

If J is a bounded non-decreasing càd-làg function defined on $\mathbb{R}$. Its generalized inverse J^{-1} is given by $J^{-1}(t) = \inf\{s \in \mathbb{R}; J(s) \geq t\}$, for $t \in \mathbb{R}$, with the convention that $\inf \emptyset = +\infty$ and $\inf \mathbb{R} = -\infty$. We have for $s, t \in \mathbb{R}$:

$$J(t) \geq s \Leftrightarrow t \geq J^{-1}(s), \quad J^{-1} \circ J(t) \leq t \quad \text{and} \quad J \circ J^{-1} \circ J(t) = J(t). \tag{3.1}$$

We define the set of points where J is increasing on their left:

$$I_g(J) = \{t \in \mathbb{R}; u < t \Leftrightarrow J(u) < J(t)\}. \tag{3.2}$$

We have:

$$\mathbf{1}_{(I_g(J))^c}\, dJ = 0 \quad \text{a.e.}, \tag{3.3}$$

$$J^{-1}(\mathbb{R}) \subset I_g(J) \cup \{\pm\infty\} \tag{3.4}$$

and for $s \in \mathbb{R}$, $t \in I_g(J)$:

$$J(t) \leq s \Leftrightarrow t \leq J^{-1}(s) \quad \text{and} \quad J^{-1} \circ J(t) = t. \tag{3.5}$$

Notice that if J is continuous in addition, then we have for $t \in J(\mathbb{R})$:

$$J \circ J^{-1}(t) = t. \tag{3.6}$$

3.2.2 Cdf and copula

Let $X = (X_1, \dots, X_d)$ be a random vector on $\mathbb{R}^d$. Its cumulative distribution function (cdf), denoted by F is defined by: $F(x) = \mathbb{P}(X \leq x)$, $x \in \mathbb{R}^d$. The corresponding one-dimensional marginals cdf are $(F_i, 1 \leq i \leq d)$ with $F_i(t) = \mathbb{P}(X_i \leq t)$, $t \in \mathbb{R}$. The cdf F is called a copula if X_i is uniform on $I = [0, 1]$ for all $1 \leq i \leq d$. (Notice a copula is characterized by its values on I^d only.)

We define $\mathcal{L}_d$ as the set of cdfs on $\mathbb{R}^d$ whose one-dimensional marginals cdfs are continuous, and $\mathcal{C} \subset \mathcal{L}_d$ as the subset of copulas. We set $\mathcal{L}_d^0$ (resp. $\mathcal{C}^0$) the subset of absolutely continuous cdf (resp. copulas) on $\mathbb{R}^d$.

Let us define for a cdf F with one-dimensional marginals $(F_i, 1 \leq i \leq d)$ the function C_F defined on I^d:

$$C_F(y) = F(F_1^{-1}(y_1), \dots, F_d^{-1}(y_d)), \quad y = (y_1, \dots, y_d) \in I^d. \tag{3.7}$$

If $F \in \mathcal{L}_d$, then C_F defined by (3.7) is a copula thanks to (3.6). According to Sklar's theorem, F is then completely characterized by its one-dimensional marginals cdf $(F_i, 1 \leq i \leq d)$ and the associated copula C_F which contains all information on the dependence:

$$F(x) = C_F\left(F_1(x_1), \dots, F_d(x_d)\right), \quad x = (x_1, \dots, x_d) \in \mathbb{R}^d. \tag{3.8}$$

Equivalently, if $X = (X_1, \dots, X_d)$ has cdf F, then C_F is the cdf of the random vector:

$$(F_1(X_1), \dots, F_d(X_d)). \tag{3.9}$$

3.2.3 Order statistics

For a d-dimensional cdf F, we write $\mathbb{P}_F$ for the distribution of a random vector $X = (X_1, \ldots, X_d)$ with cdf F. A d-dimensional cdf F is a cdf of order statistics (and we shall say that X is a vector of order statistics) if $\mathbb{P}_F(X_1 \leq X_2 \leq \ldots \leq X_d) = 1$. Let us denote by $\mathcal{L}_d^{OS} \subset \mathcal{L}_d$ the set of all cdf of order statistics with continuous one-dimensional marginals cdf. The d-tuples $(F_i, 1 \leq i \leq d)$ of marginal cdf's then verify $F_{i-1} \geq F_i$ for all $2 \leq i \leq d$. Let $\mathcal{F}_d$ be the set of d-tuples of continuous one-dimensional cdf's compatible with the marginals cdf of order statistics:

$$\mathcal{F}_d = \{\mathbf{F} = (\mathbf{F}_i, 1 \leq i \leq d) \in (\mathcal{L}_1)^d; \quad \mathbf{F}_{i-1} \geq \mathbf{F}_i, \forall 2 \leq i \leq d\}. \tag{3.10}$$

For a given $\mathbf{F} = (\mathbf{F}_i, 1 \leq i \leq d)$ in $\mathcal{F}_d$, we define the set of cdf's F of order statistics with marginals cdf $\mathbf{F}$:

$$\mathcal{L}_d^{OS}(\mathbf{F}) = \{F \in \mathcal{L}_d^{OS}; \quad F_i = \mathbf{F}_i,\ 1 \leq i \leq d\}. \tag{3.11}$$

If $\mathbf{F} \in \mathcal{F}_d$, then we have $\mathcal{L}_d^{OS}(\mathbf{F}) \neq \emptyset$, since the cdf of $(\mathbf{F}_1^{-1}(U), \ldots, \mathbf{F}_d^{-1}(U))$, U uniformly distributed on I, belongs to $\mathcal{L}_d^{OS}(\mathbf{F})$. We define $\mathcal{C}^{OS}(\mathbf{F})$ the set of copulas of order statistics with marginals $\mathbf{F}$:

$$\mathcal{C}^{OS}(\mathbf{F}) = \{C_F \in \mathcal{C}; F \in \mathcal{L}_d^{OS}(\mathbf{F})\}. \tag{3.12}$$

According to Sklar's theorem, the map $F \mapsto C_F$ is a bijection between $\mathcal{L}_d^{OS}(\mathbf{F})$ and $\mathcal{C}^{OS}(\mathbf{F})$ if $\mathbf{F} \in \mathcal{F}_d$.

3.2.4 Entropy

Let h be a reference probability density function on $\mathbb{R}$. We define $h^{\otimes d}(x) = \prod_{i=1}^d h(x_i)$ for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. The relative Shannon-entropy for a c.d.f. $F \in \mathcal{L}_d$ is given by:

$$H_h(F) = \begin{cases} -\infty & \text{if } F \in \mathcal{L}_d \setminus \mathcal{L}_d^0, \\ -\int_{\mathbb{R}^d} f \log\left(f/h^{\otimes d}\right) & \text{if } F \in \mathcal{L}_d^0, \end{cases} \tag{3.13}$$

with f the density of F. Notice that $H_h(F) \in [-\infty, 0]$ is well defined. We will use the notation $H_h(X) = H_h(F)$ if X is a random vector with cdf F and $H_h(f) = H_h(F)$ if F has density f. We shall simply write $H(F)$ (resp. $H(X)$ and $H(f)$) instead of $H_h(F)$ (resp. $H_h(X)$ and $H_h(f)$) when $h = \mathbf{1}_{[0,1]}$. Note that $H(F)$ can be finite only if F is the cdf of a probability distribution on $[0,1]^d$.

According to the next lemma, the relative entropy of any $F \in \mathcal{L}_d^{1c}$ can be decomposed into the relative entropy of the one-dimensional marginals cdf $(F_i, 1 \leq i \leq d)$ and the entropy of the associated copula C_F.

Lemma 3.1. *Let $F \in \mathcal{L}_d^{1c}$. We have:*

$$H_h(F) = H(C_F) + \sum_{i=1}^d H_h(F_i). \tag{3.14}$$

Proof. It is left to the reader to check that F has a density, say f, if and only if F_i has a density, say f_i, for $1 \leq i \leq d$ and C_F has a density, say c_F. Furthermore, in this case, we have:

$$f(x) = c_F(F_1(x_1), \ldots, F_d(x_d)) \prod_{i=1}^d f_i(x_i) \quad \text{a.e. for } x = (x_1, \ldots, x_d) \in \mathbb{R}^d,$$

as well as, with the convention $0/0 = 0$,

$$c_F(u) = \frac{f(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d))}{\prod_{i=1}^d f_i(F_i^{-1}(u_i))} \quad \text{a.e. for } u = (u_1, \ldots, u_d) \in I^d.$$

On the one hand, if F does not have a density then we have $H_h(F) = -\infty$. Since F does not have a density, then one of the F_i or C_F does not have a density either, and then $H(C_F) + \sum_{i=1}^d H_h(F_i) = -\infty$. Thus (3.14) holds.

On the other hand, let us assume that F has a density, say f. Elementary computations give with $x = (x_1, \dots, x_d)$ and $1 \le i \le d$:

$$H_h(F_i) = -\int_{\mathbb{R}} f_i(x_i) \log((f_i/h)(x_i))\, dx_i = -\int_{\mathbb{R}^d} f(x) \log((f_i/h)(x_i))\, dx.$$

We also have with $u = (u_1, \dots, u_d)$ and $x = (x_1, \dots, x_d)$:

$$\begin{aligned} H(C_F) = -\int_{[0,1]^d} c_F \log(c_F) &= -\int \frac{f(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))}{\prod_{i=1}^d f_i(F_i^{-1}(u_i))} \log\left(c_F(u)\right)\, du \\ &= -\int f(x) \log\left(c_F(F_1(x_1), \dots, F_d(x_d))\right)\, dx, \end{aligned}$$

where, for the last equality, we used the change of variable $F_i(x_i) = u_i$ (for $x_i \in I_g(F_i)$) so that $F_i^{-1}(u_i) = F_i^{-1} \circ F_i(x_i) = x_i$ holds $f_i(x_i)dx_i$-a.e and that $f(x)dx = 0$ on $\left(\otimes_{i=1}^d I_g(F_i)\right)^c$. Then use that $f(x) = c_F(F_1(x_1), \dots, F_d(x_d))\ \prod_{i=1}^d f_i(x_i)$ a.e. for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ to deduce that $H(C_F) + \sum_{i=1}^d H_h(F_i) = -\int f \log(f/h^{\otimes d}) = H_h(F)$. □

Remark 3.2. Notice that if F_i has density f_i for $1 \le i \le d$, then one can choose the reference probability density $h(t) = \frac{1}{d}\sum_{i=1}^d f_i(t)$ so that $H_h(F_i) \ge -\log(d)$. In this case, $H_h(F)$ is finite if and only if $H(C_F)$ is finite, and we have:

$$H(C_F) = H_h(F) - \sum_{i=1}^d H_h(F_i) = -\int f(x) \log\left(\frac{f(x)}{\prod_{i=1}^d f_i(x_i)}\right) dx.$$

Thus, $H(C_F)$ is the relative entropy of the cdf F with respect to the probability distribution with cdf $\otimes_{i=1}^d F_i$ of independent real valued random variables with the same one-dimensional marginal as the one with cdf F. This emphasizes the fact that $H_h(F) - \sum_{i=1}^d H_h(F_i)$, when it is well defined, does not depend on h.

For $\mathbf{F} = (\mathbf{F}_i, 1 \le i \le d) \in \mathcal{F}_d$, we define $\mathbb{J}(\mathbf{F})$ taking values in $[0, +\infty]$ by:

$$\mathbb{J}(\mathbf{F}) = \sum_{i=2}^d \int_{\mathbb{R}} \mathbf{F}_i(dt)\ \left|\log\left(\mathbf{F}_{i-1}(t) - \mathbf{F}_i(t)\right)\right|. \tag{3.15}$$

Our aim is to find the cdf $F^* \in \mathcal{L}_d^{OS}(\mathbf{F})$ which maximizes the entropy H_h. We shall see that this is possible if and only if $\mathbb{J}(\mathbf{F})$ is finite. From an information theory point of view, this is the distribution which is the least informative among distributions of order statistics with given one-dimensional marginals cdf $\mathbf{F}$. Since the vector of marginal distribution functions $\mathbf{F}$ is fixed, thanks to (3.14), we notice that $H_h(F)$ is maximal on $\mathcal{L}_d^{OS}(\mathbf{F})$ if and only if $H(C_F)$ is maximal on $\mathcal{C}^{OS}(\mathbf{F})$. Therefore we focus on finding the copula $C^* \in \mathcal{C}^{OS}(\mathbf{F})$ which maximizes the entropy H. We will give the solution of this problem in Section 3.5 under some additional hypotheses on $\mathbf{F}$.

3.3 Symmetric copulas with given order statistics

In this Section, we introduce an operator on the set $\mathcal{C}^{OS}(\mathbf{F})$ of copulas of order statistics with fixed marginals cdf $\mathbf{F}$. This operator assigns to a copula $C \in \mathcal{C}^{OS}(\mathbf{F})$ the copula of the exchangeable random vector associated to the order statistics with marginals cdf $\mathbf{F}$ and copula C. We show that this operator is a bijection between $\mathcal{C}^{OS}(\mathbf{F})$ and a set of symmetric copulas which can be characterized by their multidiagonal, which is a generalization of the well-known diagonal section of copulas. This bijection has good properties with respect to the entropy H, giving us a problem equivalent to maximizing H on $\mathcal{C}^{OS}(\mathbf{F})$. We shall solve this problem in Section 3.4.

3.3.1 Symmetric copulas

For $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ we define $x^{OS} = (x_{(1)}, \dots, x_{(d)})$ the ordered vector (increasing order) of x, where $x_{(1)} \leq \cdots \leq x_{(d)}$ and $\sum_{i=1}^d \hat{\delta}_{x_i} = \sum_{i=1}^d \hat{\delta}_{x_{(i)}}$, with $\hat{\delta}_t$ the Dirac mass at $t \in \mathbb{R}$.

Let $\mathcal{S}_d$ be the set of permutations on $\{1, \dots, d\}$. For $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ and $\pi \in \mathcal{S}_d$, we set $x_\pi = (x_{\pi(1)}, \dots, x_{\pi(d)})$. A function h defined on $\mathbb{R}^d$ is symmetric if $h(x_\pi) = h(x)$ for all $\pi \in \mathcal{S}_d$. A random vector X taking values in $\mathbb{R}^d$ is exchangeable if X_π is distributed as X for all $\pi \in \mathcal{S}_d$. In particular a random vector X taking values in $\mathbb{R}^d$ is exchangeable if and only if its cdf is symmetric. Let $\mathcal{L}_d^{sym} \subset \mathcal{L}_d$ (resp. $\mathcal{C}^{sym} \subset \mathcal{L}_d$) denote the set of symmetric cdf (resp. copulas) on $\mathbb{R}^d$.

Let $F \in \mathcal{L}_d$ and define its symmetrization $F^{sym} \in \mathcal{L}_d^{sym}$ by:

$$F^{sym}(x) = \frac{1}{d!} \sum_{\pi \in \mathcal{S}_d} F(x_\pi), \quad x \in \mathbb{R}^d. \tag{3.16}$$

In particular, if X is a random vector taking values in $\mathbb{R}^d$ with cdf F and Π is a random variable independent of X, uniformly distributed on $\mathcal{S}_d$, then X_Π is exchangeable with cdf F^{sym}.

We define the following operator on the set of copulas of order statistics.

Definition 3.3. Let $\mathbf{F} \in (\mathcal{L}_1)^d$. For $C \in \mathcal{C}$ we define $S_{\mathbf{F}}(C)$ as the copula of the exchangeable random variable X_Π, where X is a random vector on $\mathbb{R}^d$ with one-dimensional marginals cdf $\mathbf{F}$ and copula C and Π is an independent random variable uniform on $\mathcal{S}_d$.

The application $S_{\mathbf{F}}$ is well-defined on $\mathcal{C}$ and takes values in $\mathcal{C}^{sym}$. In the above definition, with $\mathbf{F} = (\mathbf{F}_i, 1 \leq i \leq d)$, the one-dimensional marginals cdf of X_Π are equal to:

$$G = \frac{1}{d} \sum_{i=1}^d \mathbf{F}_i. \tag{3.17}$$

Since the one-dimensional marginals cdf $\mathbf{F}_i$ are continuous, we get that G is continuous and thus the cdf of X_Π belongs to $\mathcal{L}_d$. In particular, thanks to Sklar's theorem, the copula of X_Π is indeed uniquely defined.

Combining (3.8), (3.7) and (3.16), we can give an explicit formula for $S_{\mathbf{F}}(C)$:

$$S_{\mathbf{F}}(C)(u) = \frac{1}{d!} \sum_{\pi \in \mathcal{S}_d} C\left(\mathbf{F}_1(G^{-1}(u_{\pi(1)})), \dots, \mathbf{F}_d(G^{-1}(u_{\pi(d)}))\right), \quad u \in I^d. \tag{3.18}$$

Remark 3.4. The copula $S_{\mathbf{F}}(C)$ is not equal in general to the exchangeable copula C^{sym} defined similarly to (3.16) by $C^{sym} = (1/d!) \sum_{\pi \in \mathcal{S}_d} C(x_\pi)$. However this is the case if the one-dimensional marginals cdf $\mathbf{F}_i$ are all equal, in which case $\mathbf{F}_i = G$ for all $1 \leq i \leq d$.

If X is a random vector on $\mathbb{R}^d$, let $X^{OS} = (X_{(1)}, \dots, X_{(d)})$ be the order statistics of X. The proof of the next Lemma is elementary.

Lemma 3.5. *Let X be a random vector on $\mathbb{R}^d$ with cdf F and Π a random variable independent of X, uniformly distributed on $\mathcal{S}_d$. We have:*
- *If $F \in \mathcal{L}_d^{OS}$, then a.s. $(X_\Pi)^{OS} = X$*
- *If $F \in \mathcal{L}_d^{sym}$, then $(X^{OS})_\Pi$ has the same distribution as X.*

For $\mathbf{F} \in \mathcal{F}_d$, we define the set of copulas $\mathcal{C}^{sym}(\mathbf{F}) \subset \mathcal{C}^{sym}$ as the image of $\mathcal{C}^{OS}(\mathbf{F})$ by the symmetrizing operator $S_{\mathbf{F}}$:

$$\mathcal{C}^{sym}(\mathbf{F}) = S_{\mathbf{F}}(\mathcal{C}^{OS}(\mathbf{F})). \tag{3.19}$$

The following Lemma is one of the main result of this section.

Lemma 3.6. *Let $\mathbf{F} \in \mathcal{F}_d$. The symmetrizing operator $S_{\mathbf{F}}$ is a bijection from $\mathcal{C}^{OS}(\mathbf{F})$ onto $\mathcal{C}^{sym}(\mathbf{F})$.*

Proof. Let $C_1, C_2 \in \mathcal{C}^{OS}(\mathbf{F})$ with $S_{\mathbf{F}}(C_1) = S_{\mathbf{F}}(C_2)$. Let X and Y be random vectors with one-dimensional marginals cdf $\mathbf{F}$ and copula C_1, C_2 respectively. Since $C_1, C_2 \in \mathcal{C}^{OS}(\mathbf{F})$, we get that X and Y are order statistics. Notice X_Π and Y_Π have the same one-dimensional marginals according to (3.17) and same copula given by $S_{\mathbf{F}}(C_1) = S_{\mathbf{F}}(C_2)$. Therefore X_Π and Y_Π have the same distribution. Thus, their corresponding order statistics $(X_\Pi)^{OS}$ and $(Y_\Pi)^{OS}$ have the same distribution. By Lemma 3.5 we get that X and Y have the same distribution as well, which implies $C_1 = C_2$. □

Remark 3.7. We have in general $\mathcal{C}^{sym}(\mathbf{F}) \neq \mathcal{C}^{OS}(\mathbf{F}) \cap \mathcal{C}^{sym}$. One exception being when the marginals cdf's $\mathbf{F}_i$ are all equal. In this case, both sides reduce to one copula which is the Fréchet-Hoeffding upper bound copula: $C^+(u) = \min u$, $u \in I^d$.

3.3.2 Multidiagonals and characterization of $\mathcal{C}^{sym}(\mathbf{F})$

Let $C \in \mathcal{C}$ be a copula and U a random vector with cdf C. The map $t \mapsto C(t, \dots, t)$ for $t \in I$, which is called the diagonal section of C, is the cdf of $\max U$. We shall consider a generalization of the diagonal section of C in the next Definition.

Definition 3.8. Let $C \in \mathcal{C}$ be a copula on $\mathbb{R}^d$ and U a random vector with cdf C. The *multidiagonal* of the copula C, $\delta_C = (\delta_{(i)}, 1 \leq i \leq d)$, is the d-tuple of the one-dimensional marginals cdf of $U^{OS} = (U_{(1)}, \dots, U_{(d)})$ the order statistics of U: for $1 \leq i \leq d$

$$\delta_{(i)}(t) = \mathbb{P}(U_{(i)} \leq t), \quad t \in I.$$

We denote by $\mathcal{D} = \{\delta_C; C \in \mathcal{C}\}$ the set of multidiagonals. Notice that $\mathcal{D} \subset \mathcal{F}_d$, see Remark 3.9. For $\delta \in \mathcal{D}$ a multidiagonal, we define $\mathcal{C}_\delta = \{C; \delta_C = \delta\}$ the set of copulas with multidiagonal δ.

A characterization of the set $\mathcal{D}$ is given by Theorem 1 of [103]: a vector of functions $\delta = (\delta_{(1)}, \dots, \delta_{(d)})$ belongs to $\mathcal{D}$ if and only if $\delta_{(i)}$ is a one-dimensional cdf and the following conditions hold:

$$\delta_{(i)} \geq \delta_{(i+1)}, \quad 1 \leq i \leq d-1, \tag{3.20}$$

$$\sum_{i=1}^{d} \delta_{(i)}(s) = ds, \qquad 0 \leq s \leq 1. \tag{3.21}$$

Remark 3.9. The condition (3.21) implies that $\delta_{(i)} \in \mathcal{L}_1$, $1 \leq i \leq d$, moreover they are d-Lipschitz. Also, it is enough to know $d-1$ functions from $\delta_{(i)}$, $1 \leq i \leq d$, the remaining one is implicitly defined by (3.21). Condition (3.20) along with the continuity of $\delta_{(i)}$ implies that any multidiagonal δ_C is compatible with the continuous marginal distributions of an order statistics, therefore $\mathcal{D} \subset \mathcal{F}_d$.

Remark 3.10. Since $\delta_{(i)}$, $1 \leq i \leq d$ are non-decreasing and d-Lipschitz, we have for almost every $t \in I$: $0 \leq (\delta_{(i)})'(t) \leq d$ and thus $\left|(\delta_{(i)})'(t) \log((\delta_{(i)})'(t))\right| \leq d \log(d)$ for $d \geq 2$. We deduce that for $d \geq 2$:

$$\left|H(\delta_{(i)})\right| \leq d \log(d). \tag{3.22}$$

Remark 3.11. Let $C \in \mathcal{C}^{sym}$ be a symmetric copula on $\mathbb{R}^d$ and U a random vector with cdf C. We check that the multidiagonal $\delta_C = (\delta_{(i)}, 1 \leq i \leq d)$ can be expressed in terms of the diagonal sections $(C_{\{i\}}, 1 \leq i \leq d)$ where for $1 \leq i \leq d$:

$$C_{\{i\}}(t) = \mathbb{P}\left(\max_{1 \leq k \leq i} U_k \leq t\right) = C(\underbrace{t, \dots, t}_{i \text{ terms}}, \underbrace{1, \dots, 1}_{d-i \text{ terms}}), \quad t \in I.$$

According to 2.8 of [103], we have for $1 \leq i \leq d$:

$$\delta_{(i)}(t) = \sum_{j=i}^{d} (-1)^{j-i} \binom{j-1}{i-1} \binom{d}{j} C_{\{j\}}(t), \quad t \in I.$$

Conversely, we can express the functions $(C_{\{i\}}, 1 \le i \le d)$ with δ_C. Let Π denote the random permutation such that $U_\Pi = U^{OS}$, where U^{OS} is the order statistics associated to U. It is well known that Π and U^{OS} are independent. Therefore, for $1 \le i \le d$ and $t \in I$, we have:

$$\begin{aligned} C_{\{i\}}(t) &= \mathbb{P}\left(\max_{1\le k\le i} U_k \le t\right) \\ &= \sum_{j=i}^{d} \mathbb{P}(U_{(j)} \le t,\ \max_{1\le k\le i} U_k = U_{(j)}) \\ &= \sum_{j=i}^{d} \mathbb{P}(U_{(j)} \le t,\ \max_{1\le k\le i} \Pi(k) = j) \\ &= \sum_{j=i}^{d} \mathbb{P}(\max_{1\le k\le i} \Pi(k) = j)\mathbb{P}(U_{(j)} \le t) \\ &= \sum_{j=i}^{d} \frac{\binom{j-1}{i-1}}{\binom{d}{i}} \delta_{(j)}(t), \end{aligned}$$

where we used the independence of Π and U^{OS} for the fourth equality, and the definition of $\delta_{(i)}$ plus the exchangeability of U for the fifth equality.

The next technical Lemma will be used in forthcoming proofs. Recall that J^{-1} denotes the generalized inverse of a non-decreasing function J, see Section 3.2.1 for its definition and properties, in particular, $J^{-1} \circ J(t) \le t$ for $t \in \mathbb{R}$. Recall also that for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, we write $G(x) = (G(x_1), \dots, G(x_d))$.

Lemma 3.12. *Let $X = (X_1, \dots, X_d)$ be a random vector on $\mathbb{R}^d$ with one-dimensional marginals cdf $(F_i, 1 \le i \le d)$. Set $G = \sum_{i=1}^d F_i/d$. We have for $1 \le i \le d$:*

$$\mathbb{P}(X_i \le G^{-1} \circ G(t)) = \mathbb{P}(X_i \le t), \quad t \in \mathbb{R}, \quad \textit{that is} \quad F_i \circ G^{-1} \circ G = F_i. \tag{3.23}$$

We also have for $x \in \mathbb{R}^d$:

$$\mathbb{P}(G(X) \le x) = \mathbb{P}(X \le G^{-1}(x)). \tag{3.24}$$

Proof. Since G is the average of the non-decreasing functions F_i, if $G(s) = G(s')$ for some $s, s' \in \mathbb{R}$, then we have $F_i(s) = F_i(s')$ for every $1 \le i \le d$. Thanks to (3.1), we have $G \circ G^{-1} \circ G(t) = G(t)$ and thus $F_i \circ G^{-1} \circ G(t) = F_i(t)$. This gives (3.23).

Recall definition (3.2) for $I_g(J)$ the set of points where the function J is increasing on their left. Since G is the average of the non-decreasing functions F_i, we deduce that $I_g(G) = \bigcup_{1\le i\le d} I_g(F_i)$. Notice that a.s. X_i belongs to $I_g(F_i)$. Thanks to (3.5), we get that a.s. $\{G(X) \le x\} = \{X \le G^{-1}(x)\}$. This gives (3.24). □

We will also require the following Lemma.

Lemma 3.13. *Let $X = (X_1, \dots, X_d)$ be a random vector on $\mathbb{R}^d$ with one-dimensional marginals cdf $(F_i, 1 \le i \le d)$. Set $G = \sum_{i=1}^d F_i/d$. We have for $1 \le i \le d$:*

$$(F_i \circ G^{-1})^{-1} = G \circ F_i^{-1}. \tag{3.25}$$

Proof. Recall Definition (3.2) for $I_g(J)$ the set of points where the function J is increasing on their left. Let $1 \le i \le d$. Thanks to (3.4), we have $F_i^{-1}(\mathbb{R}) \subset I_g(F_i) \cup \{\pm\infty\}$. Since G is the average of the non-decreasing functions F_i, we deduce that $I_g(G) = \bigcup_{1\le i\le d} I_g(F_i)$. Thus we get:

$$F_i^{-1}(\mathbb{R}) \subset I_g(G) \cup \{\pm\infty\}, \tag{3.26}$$

for all $1 \le i \le d$. The function $F_i \circ G^{-1}$ is also bounded, non-decreasing and càd-làg therefore we have for $t, s, \in \mathbb{R}$:

$$t \ge (F_i \circ G^{-1})^{-1}(s) \iff F_i \circ G^{-1}(t) \ge s \iff G^{-1}(t) \ge F_i^{-1}(s) \iff t \ge G \circ F_i^{-1}(s),$$

where we used the equivalence of (3.1) for the first and second equivalence, (3.26) and the equivalence of (3.5) for the last. This gives that $(F_i \circ G^{-1})^{-1} = G \circ F_i^{-1}$.

□

In the following Lemma, we show that for $\mathbf{F} \in \mathcal{F}_d$, all copulas in $\mathcal{C}^{sym}(\mathbf{F})$ share the same multidiagonal denoted by $\delta^{\mathbf{F}}$.

Lemma 3.14. *Let $\mathbf{F} = (\mathbf{F}_i, 1 \leq i \leq d) \in \mathcal{F}_d$. Let $C \in \mathcal{C}^{OS}(\mathbf{F})$ and U be a random vector with cdf $S_{\mathbf{F}}(C)$. Let $\delta^{\mathbf{F}} = (\delta_{(i)}, 1 \leq i \leq d)$ be the multidiagonal of $S_{\mathbf{F}}(C)$, that is the one-dimensional marginals cdf of U^{OS}, the order statistics of U. We have that $\delta^{\mathbf{F}}$ does not depend on C and for $1 \leq i \leq d$:*

$$\delta_{(i)} = \mathbf{F}_i \circ G^{-1} \quad \text{and} \quad \delta_{(i)}^{-1} = G \circ \mathbf{F}_i^{-1}, \tag{3.27}$$

with G given by (3.17). Furthermore, C is the unique copula of U^{OS}.

With obvious notation, we might simply write $\delta^{\mathbf{F}} = \mathbf{F} \circ G^{-1}$, with G given by (3.17).

Proof. Let X be a random vector of order statistics with marginals $\mathbf{F} \in \mathcal{F}_d$ and copula C. Then $S_{\mathbf{F}}(C)$ is the copula of the exchangeable random vector X_Π, where Π is uniform on $\mathcal{S}_d$ and independent of X. We have already seen in (3.17) that the one-dimensional marginals of X_Π have the same distribution given by $G \in \mathcal{L}_1$. Thanks to (3.9), we deduce that the random vector U, with cdf $S_{\mathbf{F}}(C)$, has the same distribution as $G(X_\Pi)$. Since G is non-decreasing, this implies that the order statistics of U, U^{OS}, has the same distribution as $G\left((X_\Pi)^{OS}\right)$ that is as $G(X)$, thanks to Lemma 3.5. Then use (3.24) to get for $x \in \mathbb{R}^d$:

$$\mathbb{P}(U^{OS} \leq x) = \mathbb{P}(G(X) \leq x) = \mathbb{P}(X \leq G^{-1}(x)). \tag{3.28}$$

This gives the first part of the Lemma as the multidiagonal of U is the one-dimensional marginals cdf of its order statistics. The second equation in (3.27) is due to Lemma 3.13. The fact that C is the copula of U^{OS} and its uniqueness are due to (3.28) and the continuity of $\delta_{(i)}$, see Remark 3.9. □

The next proposition shows that the set $\mathcal{C}^{sym}(\mathbf{F})$ is actually the set of symmetric copulas with diagonal section $\delta^{\mathbf{F}}$. This yields the main result of this section given by the subsequent corollary.

Proposition 3.15. *Let $\mathbf{F} \in \mathcal{F}_d$. We have $\mathcal{C}^{sym}(\mathbf{F}) = \mathcal{C}_{\delta^{\mathbf{F}}} \cap \mathcal{C}^{sym}$.*

Proof. By Lemma 3.14, we have $\mathcal{C}^{sym}(\mathbf{F}) \subset \mathcal{C}_{\delta^{\mathbf{F}}} \cap \mathcal{C}^{sym}$.

Let $C \in \mathcal{C}_{\delta^{\mathbf{F}}} \cap \mathcal{C}^{sym}$ and U be a random vector with cdf C. Let G be given by (3.17). Notice that $X = G^{-1}(U)$ is an exchangeable random vector with marginals G and copula C. Thanks to Lemma 3.5, the proof will be complete as soon as we prove that the one-dimensional marginals cdf of $X^{OS} = (X_{(1)}, \ldots, X_{(d)})$, the order statistics of X, is given by $\mathbf{F}$. Notice $X^{OS} = G^{-1}(U^{OS})$, with U^{OS} the order statistics of U whose one-dimensional marginals cdf are given by $\delta^{\mathbf{F}}$. We have for $1 \leq i \leq d$ and $t \in \mathbb{R}$:

$$\mathbb{P}(X_{(i)} \leq t) = \mathbb{P}(G^{-1}(U_{(i)}) \leq t) = \mathbb{P}(U_{(i)} \leq G(t)) = \mathbf{F}_i \circ G^{-1} \circ G(t) = \mathbf{F}_i(t),$$

where we used (3.1) for the second equality, (3.27) for the third, and (3.23) for the last. This finishes the proof. □

Corollary 3.16. *According to Proposition 3.15, (3.19) and Lemma 3.6, we get that for any $\mathbf{F} \in \mathcal{F}_d$, the symmetrizing operator $S_{\mathbf{F}}$ is a bijection between $\mathcal{C}^{OS}(\mathbf{F})$ and $\mathcal{C}_{\delta^{\mathbf{F}}} \cap \mathcal{C}^{sym}$.*

We end this Section by an ancillary result we shall use later.

Lemma 3.17. *Let $\mathbf{F} \in \mathcal{F}_d$. We have $\mathbb{J}(\mathbf{F}) = \mathbb{J}(\delta^{\mathbf{F}})$.*

Proof. Let $\mathbf{F} = (\mathbf{F}_i, 1 \le i \le d)$. We get, using (3.27) and the change of variable $s = G^{-1}(t)$ that:

$$\int_I \delta'_{(i)}(t) \left|\log\left(\delta_{(i-1)}(t) - \delta_{(i)}(t)\right)\right| dt = \int_{G^{-1}((0,1))} \mathbf{F}_i(ds) \left|\log\left(\mathbf{F}_{i-1}(s) - \mathbf{F}_i(s)\right)\right|.$$

Since $d\mathbf{F}_i = 0$ outside $G^{-1}((0,1))$ (as G is increasing as soon as $\mathbf{F}_i$ is increasing), we get that the last integration above is also over $\mathbb{R}$. We deduce that:

$$\mathbb{J}(\delta^{\mathbf{F}}) = \sum_{i=2}^{d} \int_{\mathbb{R}} \mathbf{F}_i(ds) \left|\log\left(\mathbf{F}_{i-1}(s) - \mathbf{F}_i(s)\right)\right| = \mathbb{J}(\mathbf{F}).$$

□

3.3.3 Density and entropy of copulas in $\mathcal{C}^{sym}(\mathbf{F})$

We prove in this Section that $S_{\mathbf{F}}$ preserves the absolute continuity on $\mathcal{C}^{OS}(\mathbf{F})$ for $\mathbf{F} \in \mathcal{F}_d$ and the entropy up to a constant. Let us introduce some notation. For marginals $\mathbf{F} \in \mathcal{F}_d$, let

$$\Psi_i^{\mathbf{F}} = \{s \in \mathbb{R},\ \mathbf{F}_{i-1}(s) > \mathbf{F}_i(s)\} \quad \text{for } 2 \le i \le d. \tag{3.29}$$

The complementary set $(\Psi_i^{\mathbf{F}})^c$ is the collection of the points where $\mathbf{F}_{i-1} = \mathbf{F}_i$. We define $\Sigma^{\mathbf{F}} \subset I$ as:

$$\Sigma^{\mathbf{F}} = \bigcup_{i=2}^{d} \mathbf{F}_i\left((\Psi_i^{\mathbf{F}})^c\right). \tag{3.30}$$

By Remark 3.9 we have $\mathcal{D} \subset \mathcal{F}_d$, then the definitions (3.29) and (3.30) apply for all $\delta \in \mathcal{D}$. In particular, for $\delta = (\delta_{(1)}, \dots, \delta_{(d)}) \in \mathcal{D}$ the sets Ψ_i^{δ}, $2 \le i \le d$ are open subsets of I, therefore $(\Psi_i^{\delta})^c \cap I$ is a compact subset. This and the continuity of $\delta_{(i)}$ imply that $\delta_{(i)}((\Psi_i^{\delta})^c) = \delta_{(i)}((\Psi_i^{\delta})^c \cap I)$ is also compact, hence Σ^{δ} is compact. Notice that $\{0,1\} \subset \Sigma^{\delta}$ always holds. We define $\mathcal{C}_{\delta}^0 = \mathcal{C}_{\delta} \cap \mathcal{C}^0$ the subset of absolutely continuous copulas with multidiagonal δ and the subset $\mathcal{D}^0 = \{\delta \in \mathcal{D},\ \mathcal{C}_{\delta}^0 \neq \emptyset\}$ of multidiagonals of absolutely continuous copulas. According to Theorem 2 of [103], the multidiagonal δ belongs to $\mathcal{D}^0$ if and only if it belongs to $\mathcal{D}$ and the Lebesgue measure of Σ^{δ} is zero: $\left|\Sigma^{\delta}\right| = 0$.

Lemma 3.18. *Let $\delta \in \mathcal{D}$. We have $\delta \in \mathcal{D}^0$ if and only if for all $2 \le i \le d$, a.e.:*

$$\delta'_{(i-1)} \mathbf{1}_{(\Psi_i^{\delta})^c} = \delta'_{(i)} \mathbf{1}_{(\Psi_i^{\delta})^c} = 0. \tag{3.31}$$

Furthermore, we have that $\mathbb{J}(\delta) < +\infty$ implies $\delta \in \mathcal{D}^0$.

Proof. Let J be a function defined on I, Lipschitz and non-decreasing. Let A be a Borel subset of I. We have:

$$|J(A)| = \int \mathbf{1}_{J(A)}(t)\, dt = \int_0^1 \mathbf{1}_{\{s \in J^{-1} \circ J(A)\}}\, J'(s) ds = \int_0^1 \mathbf{1}_A(s)\, J'(s) ds,$$

where we used (3.3) and (3.5) for the last equality. This gives that $|J(A)| = 0$ if and only if a.e. $J' \mathbf{1}_A = 0$. Then use that $\delta \in \mathcal{D}^0$ if and only if $\left|\delta_{(i)}((\Psi_i^{\delta})^c)\right| = 0$ for all $1 \le i \le d$ and that $\delta_{(i-1)}((\Psi_i^{\delta})^c) = \delta_{(i)}((\Psi_i^{\delta})^c)$ to conclude that $\delta \in \mathcal{D}^0$ if and only if (3.31) holds for all $2 \le i \le d$. The last part of the Lemma is clear. □

Definition 3.19. Let $\mathcal{F}_d^0 \subset \mathcal{F}_d$ be the subset of marginals $\mathbf{F}$ such that there exists an absolutely continuous cdf of order statistics with marginals $\mathbf{F}$, that is $\mathcal{L}_d^{OS}(\mathbf{F}) \cap \mathcal{L}_d^0 \neq \emptyset$.

In particular, we have $\mathcal{D}^0 \subset \mathcal{F}_d^0$. The next lemma gives a characterization of the set $\mathcal{F}_d^0$.

Lemma 3.20. *Let $\mathbf{F} \in \mathcal{F}_d$. Then $\mathbf{F} \in \mathcal{F}_d^0$ if and only if $\mathbf{F}_i \in \mathcal{L}_1^0$ for $1 \le i \le d$ and $\left|\Sigma^{\mathbf{F}}\right| = 0$. Furthermore, we have that $\mathbf{F}_i \in \mathcal{L}_1^0$ for $1 \le i \le d$ and $\mathbb{J}(\mathbf{F}) < +\infty$ imply $\mathbf{F} \in \mathcal{F}_d^0$.*

Proof. Let $F \in \mathcal{L}_d^{OS}(\mathbf{F})$. We know that $F \in \mathcal{L}_d^0$ if and only if $\mathbf{F}_i \in \mathcal{L}_1^0$ for $1 \le i \le d$ and $C_F \in \mathcal{C}^0$, the subset of absolutely continuous copulas (see for example [102]). Therefore $\mathbf{F} \in \mathcal{F}_d^0$ if and only if $\mathbf{F}_i \in \mathcal{L}_1^0$ for $1 \le i \le d$ and $\mathcal{C}^{OS}(\mathbf{F}) \cap \mathcal{C}^0 \neq \emptyset$. Recall that $\delta^{\mathbf{F}}$ is defined by (3.27). We first show that

$$\mathcal{C}^{OS}(\mathbf{F}) \cap \mathcal{C}^0 \neq \emptyset \text{ if and only if } \mathcal{C}^0_{\delta^{\mathbf{F}}} \cap \mathcal{C}^{sym} \neq \emptyset. \tag{3.32}$$

Let $C \in \mathcal{C}^{OS}(\mathbf{F}) \cap \mathcal{C}^0$. Then Lemma 3.14 ensures that $S_{\mathbf{F}}(C) \in \mathcal{C}_{\delta^{\mathbf{F}}} \cap \mathcal{C}^{sym}$. The absolute continuity of $S_{\mathbf{F}}(C)$ is a direct consequence of (3.18), (3.27) and Remark 3.9 which ensures that $\delta^{\mathbf{F}}_{(i)}, 1 \le i \le d$ are d-Lipschitz, therefore their derivatives exist a.e. on I. This ensures that $\mathcal{C}^0_{\delta^{\mathbf{F}}} \cap \mathcal{C}^{sym} \neq \emptyset$.

Conversely, let $C \in \mathcal{C}^0_{\delta^{\mathbf{F}}} \cap \mathcal{C}^{sym}$. Let U be a random vector with cdf C. Then its order statistics U^{OS} is also absolutely continuous. Therefore the copula of U^{OS}, which is $S_{\mathbf{F}}^{-1}(C)$ by Lemma 3.14, is also absolutely continuous. This proves thanks to Proposition 3.15 and Lemma 3.6 that $S_{\mathbf{F}}^{-1}(C) \in \mathcal{C}^{OS}(\mathbf{F}) \cap \mathcal{C}^0$. This gives (3.32).

Notice that $\mathcal{C}^0_{\delta^{\mathbf{F}}} \cap \mathcal{C}^{sym} \neq \emptyset$ is equivalent to $\mathcal{C}^0_{\delta^{\mathbf{F}}} \neq \emptyset$, since for any $C \in \mathcal{C}^0_{\delta^{\mathbf{F}}}$ we have that C^{sym} defined by (3.16) belongs to $\mathcal{C}^0_{\delta^{\mathbf{F}}} \cap \mathcal{C}^{sym}$. By Theorem 2 of [103], $\mathcal{C}^0_{\delta^{\mathbf{F}}} \neq \emptyset$ if and only if $\Sigma^{\delta^{\mathbf{F}}}$ has zero Lebesgue measure. The proof is then complete as one can easily verify using (3.27) that $\Sigma^{\delta^{\mathbf{F}}} = \Sigma^{\mathbf{F}}$ and thanks to Lemma 3.17. □

From now on we consider $\mathbf{F} \in \mathcal{F}_d^0$. We give an auxiliary lemma on the support of the copulas in $\mathcal{C}^{OS}(\mathbf{F}) \cap \mathcal{C}^0$.

Lemma 3.21. *Let $\mathbf{F} = (\mathbf{F}_i, 1 \le i \le d) \in \mathcal{F}_d^0$ and $C \in \mathcal{C}^{OS}(\mathbf{F}) \cap \mathcal{C}^0$. Then the density of C vanishes a.e. on $I^d \setminus T^{\mathbf{F}}$ with:*

$$T^{\mathbf{F}} = \{u = (u_1, \dots, u_d) \in I^d;\ \mathbf{F}_1^{-1}(u_1) \le \dots \le \mathbf{F}_d^{-1}(u_d)\}. \tag{3.33}$$

Proof. Let $X = (X_1, \dots, X_d)$ be a random vector of order statistics with one-dimensional marginals cdf $\mathbf{F}$ and copula $C \in \mathcal{C}^0$. Let $U = (U_1, \dots, U_d)$ be a random vector with cdf C. Then it is distributed as $(\mathbf{F}_1(X_1), \dots, \mathbf{F}_d(X_d))$, see (3.9). We get $\mathbb{P}(U \in T^{\mathbf{F}}) = 1$, since X is a vector of order statistics and $X_i \in I_g(\mathbf{F}_i)$ a.s. for $1 \le i \le d$. This gives the result. □

Now we establish the connection between the sets $\mathcal{C}^{OS}(\mathbf{F}) \cap \mathcal{C}^0$ and $\mathcal{C}^{sym}(\mathbf{F}) \cap \mathcal{C}^0$.

Lemma 3.22. *Let $\mathbf{F} \in \mathcal{F}_d^0$. The symmetrizing operator $S_{\mathbf{F}}$ is a bijection from $\mathcal{C}^{OS}(\mathbf{F}) \cap \mathcal{C}^0$ onto $\mathcal{C}^{sym}(\mathbf{F}) \cap \mathcal{C}^0$. Moreover, if $C \in \mathcal{C}^{OS}(\mathbf{F}) \cap \mathcal{C}^0$, with density function c, then the density function $s_{\mathbf{F}}(C)$ of $S_{\mathbf{F}}(C)$ is given by, for a.e. $u = (u_1, \dots, u_d) \in I^d$:*

$$s_{\mathbf{F}}(C)(u) = \frac{1}{d!}\, c\left(\delta_{(1)}(u_{(1)}), \dots, \delta_{(d)}(u_{(d)})\right) \prod_{i=1}^{d} \delta'_{(i)}(u_{(i)}). \tag{3.34}$$

Let $T^{\mathbf{F}}$ be given by (3.33). If $C \in \mathcal{C}^{sym}(\mathbf{F}) \cap \mathcal{C}_0$ with density c, then the density $s_{\mathbf{F}}^{-1}(C)$ of $S_{\mathbf{F}}^{-1}(C)$ is given by, for a.e. $u = (u_1, \dots, u_d) \in I^d$:

$$s_{\mathbf{F}}^{-1}(C)(u) = d!\, \frac{c\left(\delta_{(1)}^{-1}(u_1), \dots, \delta_{(d)}^{-1}(u_d)\right)}{\prod_{i=1}^{d} \delta'_{(i)} \circ \delta_{(i)}^{-1}(u_i)} \mathbf{1}_{T^{\mathbf{F}}}(u) \mathbf{1}_{\left\{\prod_{i=1}^{d} \delta'_{(i)} \circ \delta_{(i)}^{-1}(u_i) > 0\right\}}. \tag{3.35}$$

Proof. By Proposition 3.15, we deduce that $\mathcal{C}^{sym}(\mathbf{F}) \cap \mathcal{C}^0 = \mathcal{C}^0_{\delta^{\mathbf{F}}} \cap \mathcal{C}^{sym}$. Lemma 3.6 and the proof of Lemma 3.20 ensures that $S_{\mathbf{F}}$ is a bijection between $\mathcal{C}^{OS}(\mathbf{F}) \cap \mathcal{C}^0$ and $\mathcal{C}^{sym}(\mathbf{F}) \cap \mathcal{C}^0$. The explicit formula (3.34) can be obtained by taking the mixed derivative of the right hand side of (3.18). By Lemma 3.21, all the terms in the sum disappear except the one on the right hand side of (3.34).

To obtain (3.35), let $C \in \mathcal{C}^{sym}(\mathbf{F}) \cap \mathcal{C}^0$ with density c, and U be a random vector with cdf C. The order statistics U^{OS} derived from U is also absolutely continuous with cumulative distribution function K, and density function k given by:

$$k(u) = d!\, c(u) \mathbf{1}_{\triangle}(u), \quad u \in I^d.$$

By Lemma 3.14, $S_{\mathbf{F}}^{-1}(C)$ is the copula of U^{OS}. From (3.7), we have for $u = (u_1, \dots, u_d) \in I^d$:

$$S_{\mathbf{F}}^{-1}(C)(u) = K(\delta_{(1)}^{-1}(u_1), \dots, \delta_{(d)}^{-1}(u_d)). \tag{3.36}$$

According to (3.5), we deduce that $G^{-1} \circ G \circ \mathbf{F}_i^{-1} = \mathbf{F}_i^{-1}$ on $(0,1)$. This implies that for $s, t \in (0,1)$, $1 \le i < j \le d$:

$$\begin{aligned}
\delta_{(i)}^{-1}(s) \le \delta_{(j)}^{-1}(t) &\Leftrightarrow G \circ \mathbf{F}_i^{-1}(s) \le G \circ \mathbf{F}_j^{-1}(t) \\
&\Rightarrow G^{-1} \circ G \circ \mathbf{F}_i^{-1}(s) \le G^{-1} \circ G \circ \mathbf{F}_j^{-1}(t) \\
&\Leftrightarrow \mathbf{F}_i^{-1}(s) \le \mathbf{F}_j^{-1}(t) \\
&\Rightarrow G \circ \mathbf{F}_i^{-1}(s) \le G \circ \mathbf{F}_j^{-1}(t),
\end{aligned}$$

where we used (3.27) for the first equivalence, that G^{-1} is non-decreasing for the first implication and G is non-decreasing for the second. Thus we have, for $s, t \in (0,1)$, that the two conditions $\delta_{(i)}^{-1}(s) \le \delta_{(j)}^{-1}(t)$ and $\mathbf{F}_i^{-1}(s) \le \mathbf{F}_j^{-1}(t)$ are equivalent. Thus we deduce that the two sets

$$\left\{(u_1, \dots, u_d) \in I^d;\ \delta_{(1)}^{-1}(u_1) \le \cdots \le \delta_{(d)}^{-1}(u_d)\right\}$$

and $T^{\mathbf{F}}$ are equal up to a set of zero Lebesgue measure. Then we deduce (3.35) from (3.36). □

We give a general result on the entropy of an exchangeable random vector and the entropy of its order statistics.

Lemma 3.23. *Let X be a random vector on I^d, X^{OS} the corresponding order statistics and Π an independent uniform random variable on $\mathcal{S}_d$. Then we have:*

$$H((X^{OS})_\Pi) = \log(d!) + H(X^{OS}).$$

Proof. Let F be the cdf of X^{OS}. If $F \notin \mathcal{L}_d^0$, then the cdf F^{sym} of $(X^{OS})_\Pi$ given by (3.16) verifies also $F^{sym} \notin \mathcal{L}_d^0$, therefore $H((X^{OS})_\Pi) = H(X^{OS}) + \log(d!) = -\infty$. If $F \in \mathcal{L}_d^0$ with density function f, then the density function f^{sym} of F^{sym} is given by, for $x \in I^d$:

$$f^{sym}(x) = \frac{1}{d!} f(x^{OS}),$$

where x^{OS} is the ordered vector of x. Therefore, using that $f(x) = 0$ if $x \neq x^{OS}$, we have:

$$\begin{aligned}
H((X^{OS})_\Pi) &= -\int_{I^d} f^{sym} \log(f^{sym}) \\
&= \log(d!) - \frac{1}{d!} \int_{I^d} f(x^{OS}) \log(f(x^{OS}))\, dx \\
&= \log(d!) - \int_{I^d} f(x) \log(f(x))\, dx \\
&= \log(d!) + H(X^{OS}).
\end{aligned}$$

□

Now we are ready to give the connection between the entropy of C and $S_{\mathbf{F}}(C)$ for $C \in \mathcal{C}^{OS}(\mathbf{F})$, which is the main result of this Section. Recall the definition of $\delta^{\mathbf{F}} = (\delta_{(i)}^{\mathbf{F}}, 1 \le i \le d)$ given in Lemma 3.14 and thanks to Remark 3.10, $H(\delta_{(i)}^{\mathbf{F}})$ is finite for all $1 \le i \le d$.

Proposition 3.24. *Let $\mathbf{F} \in \mathcal{F}_d$ and $C \in \mathcal{C}^{OS}(\mathbf{F})$. Then we have:*

$$H(S_{\mathbf{F}}(C)) = \log(d!) + H(C) + \sum_{i=1}^{d} H(\delta_{(i)}^{\mathbf{F}}). \tag{3.37}$$

Proof. Let U be an exchangeable random vector with cdf $S_{\mathbf{F}}(C)$, and U^{OS} its order statistics. According to Lemma 3.14, U^{OS} has one-dimensional marginals cdf $\delta^{\mathbf{F}} = (\delta^{\mathbf{F}}_{(i)}, 1 \leq i \leq d)$ and copula C. Therefore, using Lemma 3.1 with $h = \mathbf{1}_I$, we get:

$$H(U^{OS}) = H(C) + \sum_{i=1}^{d} H(\delta^{\mathbf{F}}_{(i)}).$$

On the other hand, since $S_{\mathbf{F}}(C)$ is symmetric, Lemma 3.5 ensures that $(U^{OS})_\Pi$ has the same distribution as U. Therefore Lemma 3.23 gives:

$$H(S_{\mathbf{F}}(C)) = H(U) = H\left((U^{OS})_\Pi\right) = H(U^{OS}) + \log(d!) = H(C) + \sum_{i=1}^{d} H(\delta^{\mathbf{F}}_{(i)}) + \log(d!).$$

□

Corollary 3.25. *Since the marginals cdf* $\mathbf{F}$ *are fixed, the difference between* $H(C)$ *and* $H(S_{\mathbf{F}}(C))$ *is constant for all* $C \in \mathcal{C}^{OS}(\mathbf{F})$*. Therefore if the entropy of a copula* $C \in \mathcal{C}^{OS}(\mathbf{F})$ *is maximal, then* $S_{\mathbf{F}}(C)$ *also has maximal entropy on* $\mathcal{C}^{sym}(\mathbf{F}) = \mathcal{C}_{\delta^{\mathbf{F}}} \cap \mathcal{C}^{sym}$*.*

3.4 Maximum entropy copula with given multidiagonals

This section is a generalization of [33], where the maximum entropy copula with given diagonal section (i.e. given distribution for the maximum of its marginals) is studied.

Recall that multidiagonals of copulas on $\mathbb{R}^d$ are given by Definition 3.8. We recall some further notation: $\mathcal{D}$ denotes the set of multidiagonals; for $\delta \in \mathcal{D}$, $\mathcal{C}_\delta$ denotes the subset of copulas with multidiagonal δ; $\mathcal{C}^0$ denotes the subset copulas which are absolutely continuous, and $\mathcal{C}^0_\delta = \mathcal{C}_\delta \cap \mathcal{C}^0$. The set $\mathcal{D}^0 \subset \mathcal{D}$ contains all diagonals for which $\mathcal{C}^0_\delta \neq \emptyset$.

We give an explicit formula for C^* such that $H(C^*) = \max_{C \in \mathcal{C}_\delta} H(C)$, with H the entropy, see definition (3.13). Notice that the maximum can be taken over $\mathcal{C}^0_\delta$, since the entropy is minus infinity otherwise. When $d = 2$, the problem was solved in [33].

Let $\delta = (\delta_{(i)}, 1 \leq i \leq d) \in \mathcal{D}$ be a multidiagonal. Since $\delta_{(i)}$, $1 \leq i \leq d$ are d-Lipschitz, the entropy of $H(\delta_{(i)})$ is well defined and finite, see Remark 3.10 and $\mathbb{J}(\delta)$ given by (3.15) is also well defined and belongs to $[0, +\infty]$.

The next two lemmas provide sets on which the density of a copula with given multidiagonal is zero. For $\delta \in \mathcal{D}$, let:

$$Z_\delta = \{u \in I^d;\ \text{there exists } 1 \leq i \leq d \text{ such that } \delta'_{(i)}(u_{(i)}) = 0\}. \tag{3.38}$$

Lemma 3.26. *Let* $\delta \in \mathcal{D}^0$*. Then for all copulas* $C \in \mathcal{C}^0_\delta$ *with density* c*, we have* $c\mathbf{1}_{Z_\delta} = 0$ *a.e. that is* $c(u)\mathbf{1}_{Z_\delta}(u) = 0$ *for a.e.* $u \in I^d$*.*

Proof. By definition of $\delta_{(i)}$, we have for all $r \in I$:

$$\int_{I^d} c(u)\mathbf{1}_{\{u_{(i)} \leq r\}}\, du = \delta_{(i)}(r) = \int_0^r \delta'_{(i)}(s)\, ds.$$

This implies, by the monotone class theorem, that for all measurable subsets K of I, we have:

$$\int_{I^d} c(u)\mathbf{1}_K(u_{(i)})\, du = \int_K \delta'_{(i)}(s)\, ds.$$

Since $c \geq 0$ a.e., we deduce that a.e. $c(u)\mathbf{1}_{\{\delta'_{(i)}(u_{(i)})=0\}} = 0$ and thus a.e. $c\mathbf{1}_{Z_\delta} = 0$. □

Recall the definition of Ψ^δ_i given by (3.29) for $2 \leq i \leq d$. We also define $\Psi^\delta_1 = (0, d_1)$ with $d_1 = \inf\{s \in I; \delta_{(1)}(s) = 1\}$ and $\Psi^\delta_{d+1} = (g_{d+1}, 1)$ with $g_{d+1} = \sup\{s \in I; \delta_{(d)}(s) = 0\}$. Since Ψ^δ_i

are open subsets of I, there exist at most countably many disjoint intervals $\{(g_i^{(j)}, d_i^{(j)}),\ j \in J_i\}$ such that

$$\Psi_i^\delta = \bigcup_{j\in J_i} (g_i^{(j)}, d_i^{(j)}). \tag{3.39}$$

We denote by $m_i^{(j)} = (g_i^{(j)} + d_i^{(j)})/2$ the midpoint of these intervals for $2 \le i \le d+1$. In particular $m_{d+1} = (1 + g_{d+1})/2$. We also define $m_1 = 0$. For $\delta \in \mathcal{D}$, let:

$$L_\delta = \{u = (u_1, \dots, u_d) \in I^d; (u_{(i-1)}, u_{(i)}) \subset \Psi_i^\delta \text{ for all } 2 \le i \le d\}. \tag{3.40}$$

We have the following Lemma for all absolutely continuous copulas $C \in \mathcal{C}_\delta^0$ with density c.

Lemma 3.27. *Let $\delta \in \mathcal{D}^0$ and $2 \le i \le d$. Then for all copulas $C \in \mathcal{C}_\delta^0$ with density c, we have $c\mathbf{1}_{I\setminus L_\delta} = 0$ a.e., that is for a.e. $u = (u_1, \dots, u_d) \in I^d$, for all $s \notin \Psi_i^\delta$:*

$$c(u)\mathbf{1}_{\{u_{(i-1)}<s<u_{(i)}\}} = 0. \tag{3.41}$$

Proof. The complementary set $(\Psi_i^\delta)^c$ is given by:

$$(\Psi_i^\delta)^c = \overline{\bigcup_{j\in J_i} \{g_i^{(j)}, d_i^{(j)}\}}. \tag{3.42}$$

Let $U = (U_1, \dots, U_d)$ be a random vector with cdf $C \in \mathcal{C}_\delta^0$. For $2 \le i \le d$ and $s \in \bigcup_{j\in J_i}\{g_i^{(j)}, d_i^{(j)}\}$, that is $\delta_{(i-1)}(s) = \delta_{(i)}(s)$, we have:

$$\mathbb{P}(U_{(i-1)} < s < U_{(i)}) = \mathbb{P}(U_{(i-1)} < s) - \mathbb{P}(U_{(i)} \le s) = \delta_{(i-1)}(s) - \delta_{(i)}(s) = 0.$$

This implies that (3.41) holds a.e. for all $s \in \bigcup_{j\in J_i}\{g_i^{(j)}, d_i^{(j)}\}$. Since J_i is at most countable, we have for a.e. $u \in I^d$ and for all $s \in \bigcup_{j\in J_i}\{g_i^{(j)}, d_i^{(j)}\}$, that (3.41) holds. Since for all $u \in I$, $s \notin \Psi_i^\delta$ there exists $s' \in \bigcup_{j\in J_i}\{g_i^{(j)}, d_i^{(j)}\}$ such that

$$\mathbf{1}_{\{u_{(i-1)}<s<u_{(i)}\}} = \mathbf{1}_{\{u_{(i-1)}<s'<u_{(i)}\}},$$

we can conclude that for a.e. $u \in I^d$ and for all $s \notin \Psi_i^\delta$ (3.41) hold.

□

Notice that for all $u = (u_1, \dots, u_d) \in I^d$:

$$\mathbf{1}_{L_\delta}(u) \le \prod_{i=1}^{d} \mathbf{1}_{\Psi_i^\delta \cap \Psi_{i+1}^\delta}(u_{(i)}). \tag{3.43}$$

We define the function c_δ on I^d as, for $u = (u_1, \dots, u_d) \in I^d$:

$$c_\delta(u) = \frac{1}{d!}\,\mathbf{1}_{L_\delta}(u) \prod_{i=1}^{d} a_i(u_{(i)}), \tag{3.44}$$

where the function a_i, $1 \le i \le d$, are given by, for $t \in I$:

$$a_i(t) = K_i'(t)\,\mathrm{e}^{K_{i+1}(t)-K_i(t)}\,\mathbf{1}_{\Psi_i^\delta \cap \Psi_{i+1}^\delta}(t), \tag{3.45}$$

with for $1 \le i \le d$, $t \in (g_i^{(j)}, d_i^{(j)})$:

$$K_i(t) = \int_{m_i^{(j)}}^{t} \frac{\delta_{(i)}'(s)}{\delta_{(i-1)}(s) - \delta_{(i)}(s)}\, ds \tag{3.46}$$

and the conventions $\delta_{(0)} = 1$ and $K_{d+1} = 0$. Notice that for $t \in \Psi_1^\delta$:

$$K_1(t) = -\log(1 - \delta_{(1)}(t)). \tag{3.47}$$

Remark 3.28. The choice of $m_i^{(j)}$ for the integration lower bound in (3.46) is arbitrary as any other value in $(g_i^{(j)}, d_i^{(j)})$ would not change the definition of c_δ in (3.44).

Remark 3.29. For all $1 \le i \le d$, $j \in J_i$, $t \in (m_i^{(j)}, d_i^{(j)})$, we have the following lower bound for $K_i(t)$:

$$K_i(t) \ge \int_{m_i^{(j)}}^{t} \frac{\delta'_{(i)}(s)}{\delta_{(i-1)}(d_i^{(j)}) - \delta_{(i)}(s)}\, ds = \log\left(\frac{\delta_{(i-1)}(d_i^{(j)}) - \delta_{(i)}(m_i^{(j)})}{\delta_{(i-1)}(d_i^{(j)}) - \delta_{(i)}(t)}\right).$$

Since $\delta_{(i)}$ is non-decreasing and $\delta_{(i-1)}(d_i^{(j)}) = \delta_{(i)}(d_i^{(j)})$, we have $\lim_{t \nearrow d_i^{(j)}} K_i(t) = +\infty$.

The following proposition states that c_δ is the density of an absolutely continuous symmetric copula $C_\delta \in \mathcal{C}_\delta^0 \cap \mathcal{C}^{sym}$. It is more general than the results in [33], where only the diagonal $\delta_{(d)}$ was supposed given.

Proposition 3.30. *Let $\delta \in \mathcal{D}^0$. The function c_δ defined in (3.44)-(3.46) is the density of a symmetric copula $C_\delta \in \mathcal{C}_\delta^0 \cap \mathcal{C}^{sym}$. In addition, we have:*

$$H(C_\delta) = -\mathbb{J}(\delta) + \log(d!) + (d-1) + \sum_{i=1}^{d} H(\delta_{(i)}). \tag{3.48}$$

The proof of this proposition is given in Section 3.6.2. The following characterization of C_δ is proved in Section 3.6.6.

Proposition 3.31. *Let $\delta \in \mathcal{D}^0$. Then C_δ is the only copula in $\mathcal{C}_\delta^0$ whose density is of the form $(1/d!)\mathbf{1}_{L_\delta}(u) \prod_{i=1}^d h_i(u_{(i)})$, where h_i, $1 \le i \le d$ are measurable non-negative functions defined on I.*

The following Theorem states that the unique optimal solution of $\max_{C \in \mathcal{C}_\delta} H(C)$, if it exists, is given by C_δ. Its proof is given in Sections 3.6.7 for case (a) and 3.6.8 for case (b).

Theorem 3.32. *Let $\delta \in \mathcal{D}$.*

(a) If $\mathbb{J}(\delta) = +\infty$ then $\max_{C \in \mathcal{C}_\delta} H(C) = -\infty$.

(b) If $\mathbb{J}(\delta) < +\infty$ then $\delta \in \mathcal{D}^0$, $\max_{C \in \mathcal{C}_\delta} H(C) > -\infty$ and C_δ given in Proposition 3.30 is the unique copula such that $H(C_\delta) = \max_{C \in \mathcal{C}_\delta} H(C)$.

The copula C_δ will be called the maximum entropy copula with given multidiagonal.

Remark 3.33. In [33], we considered the problem of the maximum entropy copula with given diagonal section, that is when only $\delta_{(d)}$ is fixed. When $d = 2$, the problem considered here coincides with the problem of maximum entropy copula with given diagonal section, see Remark 3.9. For $d > 2$ the constraints of the problem discussed here are more restrictive. With the same techniques it is possible to calculate the maximum entropy copula for which the cdf of the k largest order statistics are given (that is $\delta_{(i)}$ are given for $d - k + 1 \le i \le d$). Reasoning the same way as in the proof of Theorem 3.32, we can deduce that this copula will be of the form $\prod_{i=1}^{d-k} \tilde{b}(u_{(i)}) \prod_{i=d-k+1}^{d} \tilde{a}_i(u_{(i)})$ on its domain, involving $k+1$ different functions $\tilde{b}$ and $\tilde{a}_i$, $d - k + 1 \le i \le d$ to compute based on the constraints.

3.5 Maximum entropy distribution of order statistics with given marginals

We use the results of Section 3.4 to compute the density of the maximum entropy copula for marginals $\mathbf{F} \in \mathcal{F}_d^0$ with $\mathcal{F}_d^0$ defined in Section 3.3.3. Recall $\delta^{\mathbf{F}} = (\delta_{(1)}, \dots, \delta_{(d)}) = \mathbf{F} \circ G^{-1}$ and the definition of $\Sigma^{\delta^{\mathbf{F}}}$ in (3.30). Recall K_i defined by (3.46), for $1 \le i \le d$ and $T^{\mathbf{F}}$ defined by (3.33). We define the function $c_{\mathbf{F}}$ on I^d, for $u = (u_1, \dots, u_d) \in I^d$:

$$c_{\mathbf{F}}(u) = \prod_{i=2}^{d} \frac{\mathrm{e}^{K_i(\delta_{(i-1)}^{-1}(u_{i-1})) - K_i(\delta_{(i)}^{-1}(u_i))}}{\delta_{(i-1)} \circ \delta_{(i)}^{-1}(u_i) - u_i} \mathbf{1}_{\{u \in T^{\mathbf{F}}; (\delta_{(1)}^{-1}(u_1), \dots, \delta_{(d)}^{-1}(u_d)) \in L_{\delta^{\mathbf{F}}}\}} \mathbf{1}_{\{\prod_{i=1}^{d} \delta'_{(i)} \circ \delta_{(i)}^{-1}(u_i) > 0\}}. \tag{3.49}$$

Recall the function $\mathbb{J}(\delta)$ defined on the set of multidiagonals by (3.15) and $C_{\delta^{\mathbf{F}}}$ the copula with density given by (3.44)-(3.46).

Proposition 3.34. *Let $\mathbf{F} \in \mathcal{F}_d^0$. The function $c_{\mathbf{F}}$ defined by (3.49) is the density of the copula $C_{\mathbf{F}} = \mathcal{S}_{\mathbf{F}}^{-1}(C_{\delta^{\mathbf{F}}})$ which belongs to $\mathcal{C}^{OS}(\mathbf{F})$. The entropy of $C_{\mathbf{F}}$ is given by:*

$$H(C_{\mathbf{F}}) = d - 1 - \mathbb{J}(\delta^{\mathbf{F}}). \tag{3.50}$$

Proof. Since $\mathbf{F} \in \mathcal{F}_d^0$, we have that $\delta^{\mathbf{F}} \in \mathcal{D}^0$. According to Proposition 3.30, $c_{\delta^{\mathbf{F}}}$ defined by (3.44) is the density of a symmetric copula $C_{\delta^{\mathbf{F}}}$ which belongs to $\mathcal{C}^{sym}(\mathbf{F}) \cap \mathcal{C}^0$, thanks to Proposition 3.15 and Lemma 3.22. According to Lemma 3.22, formula (3.35) we get that $c_{\mathbf{F}} = s_{\mathbf{F}}^{-1}(C_{\delta^{\mathbf{F}}})$ is therefore the density of a copula $C_{\mathbf{F}}$ which belongs to $\mathcal{C}^{OS}(\mathbf{F}) \cap \mathcal{C}^0$. Use (3.35) and (3.6) to check (3.49). To conclude, use (3.37) and (3.48) to get (3.50). □

Analogously to Lemma 3.27, we have the following restriction on the support of all $F \in \mathcal{L}_d^{OS}(\mathbf{F}) \bigcap \mathcal{L}_d^0$. Recall the definition of $\Psi_i^{\mathbf{F}}$ in (3.29). The proof of the next Lemma is similar to the proof of Lemma 3.27 and is left to the reader.

Lemma 3.35. *Let $\mathbf{F} \in \mathcal{F}_d^0$ and $t \in \mathbf{F}_i\left((\Psi_i^{\mathbf{F}})^c\right)$ for some $2 \leq i \leq d$. Then we have for all $F \in \mathcal{L}_d^{OS}(\mathbf{F}) \bigcap \mathcal{L}_d^0$ with density function f:*

$$f(x)\mathbf{1}_{\{x_{i-1}<t<x_i\}} = 0 \quad \text{for a.e. } x = (x_1, \ldots, x_d) \in S.$$

For $\delta \in \mathcal{D}$, recall the definition of L_δ in (3.40). Let $L^\delta = L_\delta \cap S$. More generally, for $\mathbf{F} \in \mathcal{F}_d$, we set:

$$L^{\mathbf{F}} = \{x = (x_1, \ldots, x_d) \in S; (x_{i-1}, x_i) \subset \Psi_i^{\mathbf{F}} \text{ for all } 2 \leq i \leq d\}. \tag{3.51}$$

The next Lemma establishes the connection between the sets $L_{\delta^{\mathbf{F}}}$ defined by (3.40) and $L^{\mathbf{F}}$.

Lemma 3.36. *Let $\mathbf{F} = (\mathbf{F}_1, \ldots, \mathbf{F}_d) \in \mathcal{F}_d^0$ with density functions $\mathbf{f}_i$ for $1 \leq i \leq d$. Let $\delta^{\mathbf{F}} = (\delta_{(1)}, \ldots, \delta_{(d)})$ given by (3.27), $T^{\mathbf{F}}$ given by (3.33) and $L_{\delta^{\mathbf{F}}}$ defined by (3.40). Then for $\prod_{i=1}^d \mathbf{f}_i(x_i)dx_1 \ldots x_d$-a.e. $x \in \mathbb{R}^d$ we have that*

$$\mathbf{1}_{T^{\mathbf{F}}}\left(\mathbf{F}_1(x_1), \ldots, \mathbf{F}_d(x_d)\right) \mathbf{1}_{L_{\delta^{\mathbf{F}}}}\left(\delta_{(1)}^{-1} \circ \mathbf{F}_1(x_1), \ldots, \delta_{(d)}^{-1} \circ \mathbf{F}_d(x_d)\right) = \mathbf{1}_{L^{\mathbf{F}}}(x)$$

Proof. According to (3.3) and (3.5), we have $\mathbf{f}_i(t)\,dt$-a.e. that $\mathbf{F}_i^{-1} \circ \mathbf{F}_i(t) = t$. This implies that $\prod_{i=1} \mathbf{f}_i(x_i)\,dx_1 \cdots dx_d$-a.e., $(\mathbf{F}_1(x_1), \ldots, \mathbf{F}_d(x_d))$ belongs to $T^{\mathbf{F}}$ if and only if $x \in S$. Recall the sets $\Psi_i^{\delta^{\mathbf{F}}}$ given by (3.29). For $\prod_{i=1} \mathbf{f}_i(x_i)\,dx_1 \cdots dx_d$-a.e. $x \in S$, we have:

$$\begin{aligned}
&(\delta_{(1)}^{-1} \circ \mathbf{F}_1(x_1), \ldots, \delta_{(d)}^{-1} \circ \mathbf{F}_d(x_d)) \in L_{\delta^{\mathbf{F}}} \\
\iff &\left(\delta_{(i-1)}^{-1} \circ \mathbf{F}_{i-1}(x_{i-1}), \delta_{(i)}^{-1} \circ \mathbf{F}_i(x_i)\right) \subset \Psi_i^{\delta^{\mathbf{F}}}, \quad 2 \leq i \leq d \\
\iff &\forall t \in \left(\delta_{(i-1)}^{-1} \circ \mathbf{F}_{i-1}(x_{i-1}), \delta_{(i)}^{-1} \circ \mathbf{F}_i(x_i)\right) : \delta_{(i-1)}(t) > \delta_{(i)}(t), \quad 2 \leq i \leq d \\
\iff &\forall t \in \left(G \circ \mathbf{F}_{i-1}^{-1} \circ \mathbf{F}_{i-1}(x_{i-1}), G \circ \mathbf{F}_i^{-1} \circ \mathbf{F}_i(x_i)\right) : \mathbf{F}_{i-1} \circ G^{-1}(t) > \mathbf{F}_i \circ G^{-1}(t), \quad 2 \leq i \leq d \\
\iff &\forall t \in (G(x_{i-1}), G(x_i)) : \mathbf{F}_{i-1} \circ G^{-1}(t) > \mathbf{F}_i \circ G^{-1}(t), \quad 2 \leq i \leq d,
\end{aligned}$$

where the first equivalence comes from the definition of $L_{\delta^{\mathbf{F}}}$, the second from the definition of $\psi_i^{\delta^{\mathbf{F}}}$, the third from (3.27) and the last from the fact that $\mathbf{f}_i(t)\,dt$-a.e. $\mathbf{F}_i^{-1} \circ \mathbf{F}_i(t) = t$. Consider the change of variable $s = G^{-1}(t)$. We have by (3.1):

$$t < G(x_i) \iff G^{-1}(t) < x_i \iff s < x_i.$$

Since $x_{i-1} \in I_g(\mathbf{F}_{i-1})$ $\mathbf{f}_{i-1}(x_{i-1})\,dx_{i-1}$-a.e., we get $x_{i-1} \in I_g(G)$ and by (3.5):

$$G(x_{i-1}) < t \iff x_{i-1} < G^{-1}(t) \iff x_{i-1} < s.$$

Therefore we deduce that $\prod_{i=1} \mathbf{f}_i(x_i)\, dx_1 \cdots dx_d$-a.e. $x \in S$:

$$\begin{aligned} &(\delta_{(1)}^{-1} \circ \mathbf{F}_1(x_1), \ldots, \delta_{(d)}^{-1} \circ \mathbf{F}_d(x_d)) \in L_{\delta^{\mathbf{F}}} \\ \Longleftrightarrow &\forall s \in (x_{i-1}, x_i) : \mathbf{F}_{i-1}(s) > \mathbf{F}_i(s), \quad 2 \leq i \leq d \\ \Longleftrightarrow &x \in L^{\mathbf{F}}. \end{aligned}$$

□

Using Proposition 3.24, we check that the copula $C_{\mathbf{F}}$ maximizes the entropy over the set $\mathcal{C}^{OS}(\mathbf{F})$. For $\mathbf{F} \in \mathcal{F}_d^0$, we define the cdf $F_{\mathbf{F}}$ as, for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$:

$$F_{\mathbf{F}}(x) = C_{\mathbf{F}}(\mathbf{F}_1(x_1), \ldots, \mathbf{F}_d(x_d)). \tag{3.52}$$

Let $\mathbf{f}_i$ denote the density function of $\mathbf{F}_i$ when it exists. Let us further note for $2 \leq i \leq d$, $t \in \mathbb{R}$:

$$\ell_i(t) = \frac{\mathbf{f}_i(t)}{\mathbf{F}_{i-1}(t) - \mathbf{F}_i(t)}. \tag{3.53}$$

When the densities $\mathbf{f}_i$ exist for all $1 \leq i \leq d$, we define the function $f_{\mathbf{F}}$ for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ as:

$$f_{\mathbf{F}}(x) = \mathbf{f}_1(x_1) \prod_{i=2}^{d} \ell_i(x_i) \exp\left(- \int_{x_{i-1}}^{x_i} \ell_i(s)\, ds \right) \mathbf{1}_{L^{\mathbf{F}}}(x), \tag{3.54}$$

with $L^{\mathbf{F}}$ given by (3.51). The next theorem asserts that the cdf $F_{\mathbf{F}}$ maximizes the entropy over the set $\mathcal{L}_d^{OS}(\mathbf{F})$ and that its density is $f_{\mathbf{F}}$. Recall $\mathbb{J}$ defined by (3.15) and h is an arbitrary probability density on $\mathbb{R}$.

Theorem 3.37. *Let $\mathbf{F} = (\mathbf{F}_i, 1 \leq i \leq d) \in \mathcal{F}_d$.*

(a) If there exists $1 \leq i \leq d$ such that $H_h(\mathbf{F}_i) = -\infty$, or if $\mathbb{J}(\mathbf{F}) = +\infty$, then we have $\max_{F \in \mathcal{L}_d^{OS}(\mathbf{F})} H_h(F) = -\infty$.

(b) If $H_h(\mathbf{F}_i) > -\infty$ for all $1 \leq i \leq d$, and $\mathbb{J}(\mathbf{F}) < +\infty$, then $\mathbf{F} \in \mathcal{F}_d^0$, $\max_{F \in \mathcal{L}_d^{OS}(\mathbf{F})} H_h(F) > -\infty$, and $F_{\mathbf{F}}$ defined in (3.52) is the unique cdf such that $H_h(F_{\mathbf{F}}) = \max_{F \in \mathcal{L}_d^{OS}(\mathbf{F})} H_h(F)$. Furthermore, the density function of $F_{\mathbf{F}}$ exists, and is given by $f_{\mathbf{F}}$ defined in (3.54). We also have:

$$H_h(F_{\mathbf{F}}) = d - 1 + \sum_{i=1}^{d} H_h(\mathbf{F}_i) - \mathbb{J}(\mathbf{F}).$$

Proof. The proof of case (a) is postponed to Section 3.6.7.

We shall assume that $H_h(\mathbf{F}_i) > -\infty$ for all $1 \leq i \leq d$ and $\mathbb{J}(\delta^{\mathbf{F}}) < +\infty$. This implies that the densities $\mathbf{f}_i$ of $\mathbf{F}_i$ exist for $1 \leq i \leq d$ and, thanks to Lemma 3.20, that $\mathbf{F} \in \mathcal{F}_d^0$. Let $F_{\mathbf{F}}$ be defined by (3.52), that is the cdf with copula $C_{\mathbf{F}}$ from Proposition 3.34 and one-dimensional marginals cdf $\mathbf{F}$. Thanks to Proposition 3.34, we have $F_{\mathbf{F}} \in \mathcal{L}_d^{OS}(\mathbf{F})$.

We deduce from (3.14), Propositions 3.15 and 3.24, Theorem 3.32 case (b) and Proposition 3.34 that $F_{\mathbf{F}}$ is the only cdf such that $H_h(F_{\mathbf{F}}) = \max_{F \in \mathcal{L}_d^{OS}(\mathbf{F})} H_h(F)$. We deduce from (3.14), (3.50) and Lemma 3.17 that:

$$H_h(F_{\mathbf{F}}) = d - 1 + \sum_{i=1}^{d} H_h(\mathbf{F}_i) - \mathbb{J}(\mathbf{F}).$$

Since the copula $C_{\mathbf{F}}$ is absolutely continuous with density $c_{\mathbf{F}}$ given in (3.49), we deduce from (3.52) that $F_{\mathbf{F}}$ has density $f_{\mathbf{F}}$ given by, for a.e. $x = (x_1, \ldots, , x_d) \in \mathbb{R}^d$:

$$f_{\mathbf{F}}(x) = c_{\mathbf{F}}\left(\mathbf{F}_1(x_1), \ldots, \mathbf{F}_d(x_d)\right) \prod_{i=1}^{d} \mathbf{f}_i(x_i). \tag{3.55}$$

Recall the expression (3.49) of $c_{\mathbf{F}}$ as well as K_i defined by (3.46), for $1 \leq i \leq d$. Using the change of variable $s = G^{-1}(t)$ and (3.23), we get (similarly to the proof of Lemma 3.17):

$$K_i \circ \delta_{(i)}^{-1} \circ \mathbf{F}_i(x_i) - K_i \circ \delta_{(i-1)}^{-1} \circ \mathbf{F}_{i-1}(x_{i-1}) = \int_{\mathbf{F}_{i-1}^{-1} \circ \mathbf{F}_{i-1}(x_{i-1})}^{\mathbf{F}_i^{-1} \circ \mathbf{F}_i(x_i)} \ell_i(s)\, ds. \tag{3.56}$$

Using (3.23), we also get:

$$\frac{\mathbf{f}_i(x_i)}{\delta_{(i-1)} \circ \delta_{(i)}^{-1} \circ \mathbf{F}_i(x_i) - \mathbf{F}_i(x_i)} = \frac{\mathbf{f}_i(x_i)}{\mathbf{F}_{i-1} \circ \mathbf{F}_i^{-1} \circ \mathbf{F}_i(x_i) - \mathbf{F}_i(x_i)}. \tag{3.57}$$

According to (3.5), we have $\mathbf{f}_i(t)\, dt$-a.e. that $\mathbf{F}_i^{-1} \circ \mathbf{F}_i(t) = t$. For $1 \leq i \leq d$, we have from (3.5) that $\prod_{i=1} \mathbf{f}_i(x_i)\, dx_1 \cdots dx_d$-a.e.:

$$\mathbf{1}_{\left\{\prod_{i=1}^d \delta'_{(i)} \circ \delta_{(i)}^{-1}(\mathbf{F}_i(x_i)) > 0\right\}} = 1. \tag{3.58}$$

We deduce from (3.49), (3.55), (3.56), (3.57), (3.58) and Lemma 3.36 that a.e. for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$:

$$f_{\mathbf{F}}(x) = \mathbf{f}_1(x_1) \prod_{i=2}^d \ell_i(x_i)\, \mathrm{e}^{-\int_{x_{i-1}}^{x_i} \ell_i(s)\, ds}\, \mathbf{1}_{L^{\mathbf{F}}}(x).$$

□

Remark 3.38. We deduce from the proof of Theorem 3.37 case (b) and Proposition 3.34, that if $\mathbf{F} = (\mathbf{F}_i, 1 \leq i \leq d) \in \mathcal{F}_d^0$, then $f_{\mathbf{F}}$ defined by (3.54) is a probability density function on $S \subset \mathbb{R}^d$.

Remark 3.39. The density $f_{\mathbf{F}}$ has a product form on $L^{\mathbf{F}}$, that is it can be written as, for a.e. $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$:

$$f_{\mathbf{F}}(x) = \prod_{i=1}^d p_i(x_i) \mathbf{1}_{L^{\mathbf{F}}}(x), \tag{3.59}$$

where the functions $(p_i, 1 \leq i \leq d)$ are measurable and non-negative.

In addition to Remark 3.39, the next Corollary asserts that $F_{\mathbf{F}}$ is the only element of $\mathcal{L}_d^{OS}(\mathbf{F})$, whose density has a product form.

Corollary 3.40. *Let $\mathbf{F} \in \mathcal{F}_d^0$. Let $F \in \mathcal{L}_d^{OS}(\mathbf{F})$ be an absolutely continuous cdf with density f given by, a.e. for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$: $f(x) = \prod_{i=1}^d h_i(x_i) \mathbf{1}_{L^{\mathbf{F}}}(x)$, with h_i, $1 \leq i \leq d$ some measurable non-negative functions on $\mathbb{R}$. Then we have $F = F_{\mathbf{F}}$ on $\mathbb{R}^d$.*

Proof. Let $X = (X_1, \ldots, X_d)$ be an order statistic with cdf F, and F^{sym} the cdf of X_Π given by (3.16), with Π uniform on $\mathcal{S}_d$ and independent of X. Then the cdf F^{sym} is also absolutely continuous, and its density f^{sym} is given by :

$$f^{sym}(x) = \frac{1}{d!} \prod_{i=1}^d h_i(x_{(i)}) \mathbf{1}_{L^{\mathbf{F}}}(x^{OS}), \tag{3.60}$$

where x^{OS} is the ordered vector of x. The one-dimensional marginal cdf's of X_Π are all equal to G given by (3.17). Let $C \in \mathcal{C}^{OS}(\mathbf{F}) \cap \mathcal{C}^0$ denote the copula of F. Then according to (3.7), the copula $S_{\mathbf{F}}(C)$ of X_Π is given by, for a.e. $u = (u_1, \ldots, u_d) \in I^d$:

$$S_{\mathbf{F}}(C)(u) = F^{sym}\left(G^{-1}(u)\right).$$

Therefore its density $s_{\mathbf{F}}(C)$ can be expressed as:

$$s_{\mathbf{F}}(C)(u) = \frac{f^{sym}(G^{-1}(u))}{\prod_{i=1}^d g\circ G^{-1}(u_i)} \prod_{i=1}^d \mathbf{1}_{\{g\circ G^{-1}(u_i)>0\}}$$
$$= \frac{1}{d!}\,\mathbf{1}_{L^{\mathbf{F}}}(G^{-1}(u^{OS})) \prod_{i=1}^d \frac{h_i\circ G^{-1}(u_{(i)})}{g\circ G^{-1}(u_{(i)})} \mathbf{1}_{\{g\circ G^{-1}(u_{(i)})>0\}},$$

where g is the density of G. Notice that for $x=(x_1,\dots,x_d)\in S$, we have a.e.:

$$\mathbf{1}_{\{\prod_{i=1}^d h_i(x_i)>0\}} \le \mathbf{1}_{\{\prod_{i=1}^d f_i(x_i)>0\}},$$

with f_i the density of X_i. Therefore by Lemma 3.36 and since G is continuous, we have that $\prod_{i=1}^d h_i\circ G^{-1}(u_{(i)})\,du_1\dots du_d$-a.e.:

$$\mathbf{1}_{L^{\mathbf{F}}}(G^{-1}(u^{OS})) = \mathbf{1}_{L_{\delta^{\mathbf{F}}}}(\delta_{(1)}^{-1}\circ\delta_{(1)}(u_{(1)}),\dots,\delta_{(d)}^{-1}\circ\delta_{(d)}(u_{(d)})).$$

By Lemma 3.14, $S_{\mathbf{F}}(C)$ belongs to $\mathcal{C}^0_{\delta^{\mathbf{F}}}$ and thus $s_{\mathbf{F}}(C)=0$ a.e. on $Z_{\delta^{\mathbf{F}}}$ defined by (3.38). Then use (3.3) and (3.5) to get that $\delta_{(i)}^{-1}\circ\delta_{(i)}(u_{(i)})=u_{(i)}$ a.e. on $Z^c_{\delta^{\mathbf{F}}}$. This gives:

$$\mathbf{1}_{L^{\mathbf{F}}}(G^{-1}(u^{OS})) = \mathbf{1}_{L_{\delta^{\mathbf{F}}}}(u^{OS}) = \mathbf{1}_{L_{\delta^{\mathbf{F}}}}(u)$$

that is $s_{\mathbf{F}}(C)$ is of the form $s_{\mathbf{F}}(C)(u) = (1/d!)\mathbf{1}_{L_{\delta^{\mathbf{F}}}}(u)\prod_{i=1}^d \bar{h}_i(u_{(i)})$ for some measurable non-negative functions $(\bar{h}_i, 1\le i\le d)$. Then, thanks to Proposition 3.31, we get that $S_{\mathbf{F}}(C)=C_{\delta^{\mathbf{F}}}$. Then, use Proposition 3.34 to get that $F=F_{\mathbf{F}}$. □

Example 3.41. We consider the following example. Let $+\infty > \lambda_1 > \dots > \lambda_d > 0$ and for $1\le i\le d$ let $\mathbf{F}_i$ be the cdf of the exponential distribution with mean $1/\lambda_i$ and density $\mathbf{f}_i(t)=\lambda_i\,\mathrm{e}^{-\lambda_i t}\,\mathbf{1}_{\{t>0\}}$. Notice that $\mathbf{F}_{i-1}>\mathbf{F}_i$ on $(0,+\infty)$, so that $L^{\mathbf{F}}=\{(x_1,\dots,x_d)\in\mathbb{R}^d; 0\le x_1\le\dots\le x_d\}$. It is easy to check that $\mathbb{J}(\mathbf{F})<+\infty$ with $\mathbf{F}=(\mathbf{F}_i, 1\le i\le d)$. Elementary computations yield that the maximum entropy density of the order statistic $(X_1,\dots,X_d)$, where X_i has distribution $\mathbf{F}_i$, is given by:

$$f_{\mathbf{F}}(x_1,\dots,x_d) = \mathbf{1}_{L^{\mathbf{F}}}(x)\,\lambda_1\,\mathrm{e}^{-\Delta_2 x_1}\left(1-\mathrm{e}^{-\Delta_2 x_1}\right)^{\lambda_2/\Delta_2}\prod_{i=2}^d \lambda_i\,\mathrm{e}^{-\Delta_{i+1}x_i}\,\frac{\left(1-\mathrm{e}^{-\Delta_{i+1}x_i}\right)^{\lambda_{i+1}/\Delta_{i+1}}}{\left(1-\mathrm{e}^{-\Delta_i x_i}\right)^{\lambda_{i-1}/\Delta_i}},$$

where $\Delta_i=\lambda_{i-1}-\lambda_i$ for $1\le i\le d+1$ and $\lambda_{d+1}=0$.

In the particular case $\lambda_i=(d-i+1)\lambda$ for some $\lambda>0$, we get:

$$f_{\mathbf{F}}(x_1,\dots,x_d) = \mathbf{1}_{L^{\mathbf{F}}}(x)\,d!\,\lambda^d\,\mathrm{e}^{-\lambda x_1}(1-\mathrm{e}^{-\lambda x_1})^{d-1}\prod_{i=2}^d \frac{\mathrm{e}^{-\lambda x_i}}{(1-\mathrm{e}^{-\lambda x_i})^2}.$$

By considering the change of variable $u_i=1-\mathrm{e}^{-\lambda x_i}$, we get the following result. For $1\le i\le d$ let $\mathbf{F}_i$ be the cdf of the $\beta(1,d-i+1)$ distribution with density $\mathbf{f}_i(t)=(d-i+1)(1-t)^{d-i}\mathbf{1}_{(0,1)}(t)$. Notice that $\mathbf{F}_{i-1}>\mathbf{F}_i$ on $(0,1)$. The maximum entropy density of the order statistic $(U_1,\dots,U_d)$, where U_i has distribution $\mathbf{F}_i$, is given by:

$$f_{\mathbf{F}}(u_1,\dots,u_d) = \mathbf{1}_{\{0<u_1<\dots<u_d<1\}}\,d!\,u_1^{d-1}\prod_{i=2}^d \frac{1}{u_i^2}.$$

Elementary computations give $H(F_{\mathbf{F}}) = -\log(d!)+2d-(d+1)\sum_{i=1}^d(1/i)$.

3.6 Proofs

3.6.1 Preliminary notations for the optimization problem

Recall notations from Sections 3.2 and 3.3. In particular if $u = (u_1, \dots, u_d) \in I^d$ then $u^{OS} = (u_{(1)}, \dots, u_{(d)})$ denote the ordered vector of u.

In order to apply the technique established in [25], we introduce the linear functional $\mathcal{A} = (\mathcal{A}_i, 1 \le i \le 2d) : L^1(I^d) \to L^1(I)^{2d}$ as, for $f \in L^1(I^d)$ and $r \in I$:

$$\mathcal{A}_i(f)(r) = \int_{I^d} f(u)\mathbf{1}_{\{u_i \le r\}}\, du \quad \text{and} \quad \mathcal{A}_{d+i}(f)(r) = \int_{I^d} f(u)\mathbf{1}_{\{u_{(i)} \le r\}}\, du \quad \text{for} \quad 1 \le i \le d.$$

Let $\delta = (\delta_{(i)}, 1 \le i \le d) \in \mathcal{D}^0$ be a multidiagonal, see Definition 3.8. We set $b^\delta = (b_i, 1 \le i \le 2d)$ given by $b_i = \mathrm{id}_I$ the identity function on I and $b_{d+i} = \delta_{(i)}$, for $1 \le i \le d$. If, for $c \in L^1(I^d)$, we have $\mathcal{A}_i(c) = b_i$, $1 \le i \le d$ and $c \ge 0$ a.e., then we deduce that c is the density of an absolutely continuous copula, say C. If we further have $\mathcal{A}_{d+i}(c) = b_{d+i}$, for $1 \le i \le d$, then δ is the multidiagonal of C.

Lemma 3.42. *Let $\delta \in \mathcal{D}^0$ and $b^\delta = (b_i, 1 \le i \le 2d)$. If $c \in L^1(I^d)$ is non-negative, symmetric and satisfies $\mathcal{A}_{d+i}(c) = b_{d+i}$ for $1 \le i \le d$, then c is the density of a copula with multidiagonal δ.*

Proof. The symmetry and non-negativity of c as well as the condition $\mathcal{A}_{d+1}(c)(1) = \int_{I^d} c = b_{d+1}(1) = 1$ ensures that c is a density function of an exchangeable random vector $V = (V_1, \dots, V_d)$ on I^d. Recall $V^{OS} = (V_{(1)}, \dots, V_{(d)})$ denotes the corresponding order statistics. By symmetry, the lemma is proved as soon as we check that $\mathcal{A}_1(c) = b_1$. We have for $r \in I$:

$$\mathcal{A}_1(c)(r) = \mathbb{P}(V_1 \le r) = \sum_{i=1}^{d} \mathbb{P}(V_{(i)} \le r | V_1 = V_{(i)}) \mathbb{P}(V_1 = V_{(i)}) = \sum_{i=1}^{d} \delta_{(i)}(r)\, \frac{1}{d} = r,$$

where we used the exchangeability of V and the definition of $\delta_{(i)}$ for the third equality, and (3.21) for the last. This gives $\mathcal{A}_1(c) = b_1$. □

3.6.2 Proof of Proposition 3.30

Let $\delta \in \mathcal{D}^0$. Lemma 3.18 implies that $\delta_{(i)}((\Psi_i^\delta)^c)$ has zero Lebesgue measure for all $2 \le i \le d$ with Ψ_i^δ given by (3.29). By construction, the function c_δ defined by (3.44) is non-negative, symmetric and well defined a.e. on I^d. Recall the notation $(g_i^{(j)}, d_i^{(j)})$ used in (3.39) and the definition (3.45) of the functions a_i. We define the functions B_i on I as, for $1 \le i \le d+1$, $t \in (g_i^{(j)}, d_i^{(j)})$ (with the conventions $\Psi_1^\delta = (0, d_1), \Psi_{d+1}^\delta = (g_{d+1}, 1)$):

$$B_{d+1}(t) = 1 \quad \text{and} \quad B_i(t) = \int_t^{d_i^{(j)}} a_i(s) B_{i+1}(s)\, ds \quad \text{for} \quad 1 \le i \le d. \tag{3.61}$$

For $t \in (\Psi_i^\delta)^c$, we set $B_i(t) = 0$. Recall K_i defined in (3.46) for $1 \le i \le d+1$ with the convention $K_{d+1} = 0$. We show that B_i can be simply expressed by K_i on Ψ_i^δ.

Lemma 3.43. *Let $1 \le i \le d+1$ and $t \in \Psi_i^\delta$. Then we have:*

$$B_i(t) = \exp\left(-K_i(t)\right). \tag{3.62}$$

Proof. For $i = d+1$, the result is trivial. We proceed by induction on i. We suppose that

$B_{i+1}(t) = \exp(-K_{i+1}(t))$ holds for some $1 \le i \le d$, and all $t \in \psi^\delta_{i+1}$. We have for $t \in (g_i^{(j)}, d_i^{(j)})$:

$$\begin{aligned} B_i(t) &= \int_t^{d_i^{(j)}} a_i(s) B_{i+1}(s)\, ds \\ &= \int_t^{d_i^{(j)}} K_i'(s)\, \mathrm{e}^{K_{i+1}(s)-K_i(s)}\, \mathbf{1}_{\Psi_i^\delta \cap \Psi_{i+1}^\delta}(s) B_{i+1}(s)\, ds \\ &= \int_t^{d_i^{(j)}} K_i'(s)\, \mathrm{e}^{-K_i(s)}\, \mathbf{1}_{\Psi_i^\delta \cap \Psi_{i+1}^\delta}(s)\, ds. \\ &= \int_t^{d_i^{(j)}} K_i'(s)\, \mathrm{e}^{-K_i(s)}\, ds \\ &= \exp(-K_i(t)), \end{aligned}$$

where we used the definition of a_i given by (3.45) for the second equality, the induction hypothesis for the third equality, $(t, d_i^{(j)}) \subset \Psi_i^\delta$ and Lemma 3.18 for the fourth equality, and finally Remark 3.29 for the fifth equality. This ends the induction. □

Similarly, we define the functions E_i on I as, for $0 \le i \le d$ as for $t \in (g_{i+1}^{(j)}, d_{i+1}^{(j)})$:

$$E_0(t) = 1, \quad \text{and} \quad E_i(t) = \int_{g_{i+1}^{(j)}}^t a_i(s) E_{i-1}(s)\, ds \quad \text{for} \quad 1 \le i \le d. \tag{3.63}$$

For $t \in (\Psi_{i+1}^\delta)^c$ we set $E_i(t) = 0$. The next Lemma gives a simple formula for E_i on Ψ_{i+1}^δ.

Lemma 3.44. *Let $0 \le i \le d$ and $t \in \Psi_{i+1}^\delta$. Then we have:*

$$E_i(t) = \left(\delta_{(i)}(t) - \delta_{(i+1)}(t)\right) \exp\left(K_{i+1}(t)\right). \tag{3.64}$$

Proof. For $i = 0$ the result is clear thanks to the convention $\delta_{(0)} = 1$ and (3.47). We proceed by induction on i. We suppose that $E_{i-1}(t) = (\delta_{(i-1)}(t) - \delta_{(i)}(t)) \exp(K_i(t))$ holds for some $1 \le i \le d$, and all $t \in \psi_i^\delta$. Let us denote $h_i = \delta_{(i-1)} - \delta_{(i)}$. Before computing $E_i(t)$ for $t \in (g_{i+1}^{(j)}, d_{i+1}^{(j)})$, we give an alternative expression for $\exp(K_i(s))$ for $s \in (g_{i+1}^{(j)}, t)$:

$$\begin{aligned} \mathrm{e}^{K_{i+1}(s)} &= \exp\left(- \int_{m_{i+1}^{(j)}}^t \frac{h_{i+1}'(u)}{h_{i+1}(u)} + \int_s^t \frac{h_{i+1}'(u)}{h_{i+1}(u)} + \int_{m_{i+1}^{(j)}}^s \frac{\delta_{(i)}'(u)}{h_{i+1}(u)}\, du \right) \\ &= \frac{h_{i+1}(t)}{h_{i+1}(s)} \exp\left(- \int_{m_{i+1}^{(j)}}^t \frac{h_{i+1}'(u)}{h_{i+1}(u)} + \int_{m_{i+1}^{(j)}}^s \frac{\delta_{(i)}'(u)}{h_{i+1}(u)}\, du \right). \end{aligned} \tag{3.65}$$

Then we have for $t \in (g_{i+1}^{(j)}, d_{i+1}^{(j)})$:

$$\begin{aligned} E_i(t) &= \int_{g_{i+1}^{(j)}}^t a_i(s) E_{i-1}(s)\, ds \\ &= \int_{g_{i+1}^{(j)}}^t K_i'(s)\, \mathrm{e}^{K_{i+1}(s)-K_i(s)}\, \mathbf{1}_{\Psi_i^\delta \cap \Psi_{i+1}^\delta}(s) E_{i-1}(s)\, ds \\ &= \int_{g_{i+1}^{(j)}}^t K_i'(s)\, \mathrm{e}^{K_{i+1}(s)}\, h_i(s) \mathbf{1}_{\Psi_i^\delta \cap \Psi_{i+1}^\delta}(s)\, ds \\ &= \int_{g_{i+1}^{(j)}}^t \delta_{(i)}'(s)\, \mathrm{e}^{K_{i+1}(s)}\, \mathbf{1}_{\Psi_i^\delta \cap \Psi_{i+1}^\delta}(s)\, ds \\ &= h_{i+1}(t) \exp\left(- \int_{m_{i+1}^{(j)}}^t \frac{h_{i+1}'(u)}{h_{i+1}(u)}\, du \right) \int_{g_{i+1}^{(j)}}^t \left(\frac{\delta_{(i)}'(s)}{h_{i+1}(s)} \exp\left(\int_{m_{i+1}^{(j)}}^s \frac{\delta_{(i)}'(u)}{h_{i+1}(u)}\, du \right) \right) ds \\ &= h_{i+1}(t) \exp\left(- \int_{m_{i+1}^{(j)}}^t \frac{h_{i+1}'(u) - \delta_{(i)}'(u)}{h_{i+1}(u)}\, du \right) \\ &= h_{i+1}(t) \exp(K_{i+1}(t)), \end{aligned}$$

where we used the definition of a_i given by (3.45) for the second equality, the induction hypothesis for the third equality, Lemma 3.18 and (3.65) for the fifth equality, and for the seventh equality we use that, for $t \in (g_{i+1}^{(j)}, m_{i+1}^{(j)})$ (similarly to Remark 3.29):

$$\int_{m_{i+1}^{(j)}}^{t} \frac{\delta'_{(i)}(s)}{h_{i+1}(s)}\,ds \le \int_{m_{i+1}^{(j)}}^{t} \frac{\delta'_{(i)}(s)}{\delta_{(i)}(s) - \delta_{(i+1)}(g_{i+1}^{(j)})}\,ds = \log\left(\frac{\delta_{(i)}(t) - \delta_{(i+1)}(g_{i+1}^{(j)})}{\delta_{(i)}(m_{i+1}^{(j)}) - \delta_{(i+1)}(g_{i+1}^{(j)})}\right),$$

giving $\lim_{t \searrow g_{i+1}^{(j)}} \int_{m_{i+1}^{(j)}}^{t} \frac{\delta'_{(i)}(s)}{h_{i+1}(s)}\,ds = -\infty$.

□

The following Lemma justifies the introduction of the functions B_i, E_i.

Lemma 3.45. *We have with $u_{(0)} = 0$ for $1 \le i \le d$, $t \in \Psi_i^\delta$:*

$$\int_{I^d} c_\delta(u) \mathbf{1}_{\{u_{(i-1)} \le t \le u_{(i)}\}}\,du = B_i(t) E_{i-1}(t). \tag{3.66}$$

Proof. The definition (3.61) of B_i for $1 \le i \le d$ gives that for $t \in I$:

$$B_i(t) = \int a_i(r_i) a_{i+1}(r_{i+1}) \dots a_d(r_d) \mathbf{1}_{\{t \le r_i \le r_{i+1} \le \dots \le r_d \le 1\}} \mathbf{1}_{\{[t, r_i) \subset \Psi_i^\delta\}} \prod_{j=i}^{d-1} \mathbf{1}_{\{(r_j, r_{j+1}) \subset \Psi_{j+1}^\delta\}}\,dr,$$

with $r = (r_i, r_{i+1}, \dots, r_d) \in I^{d-i+1}$. Similarly, we have for $1 \le i \le d$, $t \in I$ that $E_{i-1}(t)$ is equal to:

$$\int a_1(q_1) a_2(q_2) \dots a_{i-1}(q_{i-1}) \mathbf{1}_{\{0 \le q_1 \le q_2 \le \dots \le q_{i-1} \le t\}} \mathbf{1}_{\{(q_{i-1}, t] \subset \Psi_i^\delta\}} \prod_{j=1}^{i-2} \mathbf{1}_{\{(q_j, q_{j+1}) \subset \Psi_{j+1}^\delta\}}\,dq,$$

with $q = (q_1, q_2, \dots, q_{i-1}) \in I^{i-1}$. Multiplying $B_i(t)$ with $E_{i-1}(t)$ gives:

$$\begin{aligned}
B_i(t) E_{i-1}(t) &= \int_\triangle \prod_{j=1}^{d} a_j(u_j) \mathbf{1}_{\{u_{i-1} \le t \le u_i\}} \prod_{j=1}^{d-1} \mathbf{1}_{\{(u_j, u_{j+1}) \subset \Psi_{j+1}^\delta\}}\,du \\
&= \int_\triangle \prod_{j=1}^{d} a_j(u_j) \mathbf{1}_{\{u_{i-1} \le t \le u_i\}} \mathbf{1}_{L^\delta}(u)\,du \\
&= d! \int_\triangle c_\delta(u) \mathbf{1}_{\{u_{i-1} \le t \le u_i\}}\,du \\
&= \int_{I^d} c_\delta(u) \mathbf{1}_{\{u_{(i-1)} \le t \le u_{(i)}\}}\,du,
\end{aligned}$$

where we used the symmetry of c_δ for the fourth equality. □

Lemma 3.45 with $i = 1$ ensures that $\int_{I^d} c_\delta(u)\,du = \lim_{t \searrow 0} B_1(t) E_0(t) = 1$, that is c_δ a probability density function on I^d. Now we compute $\mathcal{A}_{d+1}(c_\delta)$. We have, for $t \in \Psi_i^\delta$:

$$\mathcal{A}_{d+1}(c_\delta)(t) = \int_{I^d} c_\delta(u) \mathbf{1}_{\{u_{(1)} \le t\}}\,du = 1 - \int_{I^d} c_\delta(u) \mathbf{1}_{\{u_{(1)} \ge t\}}\,du = 1 - B_1(t) E_0(t) = \delta_{(1)}(t),$$

where we used Lemma 3.45 with $i = 1$ for the third equality, then (3.62) and (3.47) for the fourth equality. By continuity this gives $\mathcal{A}_{d+1}(c_\delta) = \delta_{(1)}$ on I. For $2 \le i \le d$, we have by induction for $t \in \Psi_i^\delta$:

$$\begin{aligned}
\mathcal{A}_{d+i}(c_\delta)(t) &= \int_{I^d} c_\delta(u) \mathbf{1}_{\{u_{(i)} \le t\}}\,du \\
&= \int_{I^d} c_\delta(u) \mathbf{1}_{\{u_{(i-1)} \le t\}}\,du - \int_{I^d} c_\delta(u) \mathbf{1}_{\{u_{(i-1)} \le t \le u_{(i)}\}}\,du \\
&= \mathcal{A}_{d+i-1}(c_\delta)(t) - B_i(t) E_{i-1}(t) \\
&= \delta_{(i-1)}(t) - \big(\delta_{(i-1)}(t) - \delta_{(i)}(t)\big) \\
&= \delta_{(i)}(t),
\end{aligned}$$

where we used the induction and Lemma 3.45 for the third equality, as well as (3.62) and (3.64) for the fourth. By continuity, we obtain $\mathcal{A}_{d+i}(c_\delta) = \delta_{(i)}$ on I. Then use Lemma 3.42 to get that c_δ is the density of a (symmetric) copula, say C_δ, with multidiagonal δ.

To conclude, we compute the entropy $H(C_\delta) = -\int_{I^d} c_\delta \log(c_\delta)$.

Lemma 3.46. *We have:*

$$H(C_\delta) = \log(d!) + \sum_{i=1}^{d} H(\delta_{(i)}) + (d-1) - \mathbb{J}(\delta).$$

Proof. Recall that for $u \in L_\delta$:

$$\log(c_\delta(u)) = -\log(d!) + \sum_{i=1}^{d} \log\left(\delta'_{(i)}(u_{(i)})\right) - \sum_{i=2}^{d} \log\left(\delta_{(i-1)}(u_{(i)}) - \delta_{(i)}(u_{(i)})\right) - \sum_{i=2}^{d} \left(K_i(u_{(i)}) - K_i(u_{(i-1)})\right),$$

where we used (3.47) to express $a_1 = \delta'_{(1)}\, \mathrm{e}^{K_2}$ a.e., so that the sums in the last two terms start at $i = 2$.

We first show that the function $u \mapsto c_\delta(u) \log(\delta'_{(i)}(u_{(i)}))$ belongs to $L^1(I^d)$ for all $1 \le i \le d$. Since $\mathcal{A}_{d+i}(c_\delta) = \delta_{(i)}$, we deduce that for $1 \le i \le d$ and any measurable non-negative function h defined on I:

$$\int_{I^d} c_\delta(u) h(u_{(i)})\, du = \int_I \delta'_{(i)}(t) h(t)\, dt. \tag{3.67}$$

In particular, we get:

$$\int_{I^d} c_\delta(u) \left|\log(\delta'_{(i)}(u_{(i)}))\right| \, du = \int_I \delta'_{(i)}(t) \left|\log(\delta'_{(i)}(t))\right| \, dt,$$

which is finite thanks to Remark 3.10. Therefore the function $u \mapsto c_\delta(u) \log(\delta'_{(i)}(u_{(i)}))$ is indeed in $L^1(I^d)$, and its integral $J_{1,i}$ is given by:

$$J_{1,i} = \int_{I^d} c_\delta(u) \log(\delta'_{(i)}(u_{(i)}))\, du = \int_I \delta'_{(i)}(t) \log(\delta'_{(i)}(t))\, dt = -H(\delta_{(i)}).$$

We proceed by showing that $u \mapsto c_\delta(u)(K_i(u_{(i)}) - K_i(u_{(i-1)}))$ belongs to $L^1(I^d)$ for $2 \le i \le d$. Since this is a non-negative function, a direct calculation of its integral $J_{2,i}$ gives:

$$\begin{aligned} J_{2,i} &= d! \int_{\triangle} c_\delta(u) \left(K_i(u_i) - K_i(u_{i-1})\right) du, \\ &= \int_{I^2} (E_{i-2} a_{i-1})(u_{i-1}) \left(K_i(u_i) - K_i(u_{i-1})\right) (a_i B_{i+1})(u_i) \mathbf{1}_{\{u_{i-1} \le u_i, (u_{i-1}, u_i) \subset \Psi_i^\delta\}}\, du_{i-1} du_i, \end{aligned}$$

where we used the symmetry of c_δ for the first equality; the definition of the functions B_i and E_i given by (3.61) and (3.63) for the second equality. Using (3.45),(3.62), (3.64) and Lemma 3.18, we have:

$$E_{i-2} a_{i-1} = \delta'_{(i-1)}\, \mathrm{e}^{K_i}\, \mathbf{1}_{\Psi_{i-1}^\delta \cap \Psi_i^\delta} = \delta'_{(i-1)}\, \mathrm{e}^{K_i} \quad \text{and} \quad a_i B_{i+1} = K_i'\, \mathrm{e}^{-K_i}\, \mathbf{1}_{\Psi_i^\delta \cap \Psi_{i+1}^\delta} = K_i'\, \mathrm{e}^{-K_i}.$$

Therefore we have:

$$\begin{aligned} J_{2,i} &= \sum_{j \in J_i} \int_{g_i^{(j)}}^{d_i^{(j)}} \delta'_{(i-1)}(u_{i-1}) \left(\int_{u_{i-1}}^{d_i^{(j)}} K_i'(u_i) \left(K_i(u_i) - K_i(u_{i-1})\right) \mathrm{e}^{K_i(u_{i-1}) - K_i(u_i)}\, du_i \right) du_{i-1} \\ &= \sum_{j \in J_i} \int_{g_i^{(j)}}^{d_i^{(j)}} \delta'_{(i-1)}(u_{i-1}) \left(\int_0^{+\infty} s\, \mathrm{e}^{-s}\, ds \right) du_{i-1} \\ &= \sum_{j \in J_i} \int_{g_i^{(j)}}^{d_i^{(j)}} \delta'_{(i-1)}(u_{i-1})\, du_{i-1} \\ &= 1, \end{aligned}$$

where we applied the change of variable $s = K_i(u_i) - K_i(u_{i-1})$ and Remark 3.29 for the fourth equality; finally Lemma 3.18 for the sixth equality.

Let us define $J_{3,i}$, $2 \leq i \leq d$ as:

$$J_{3,i} = -\int_{I^d} c_\delta(u) \log(\delta_{(i-1)}(u_{(i)}) - \delta_{(i)}(u_{(i)}))\, du.$$

Notice that the integrand is non-positive a.e., since for $t \in I$, $2 \leq i \leq d$, we have $\delta_{(i-1)}(t) - \delta_{(i)}(t) \leq 1$. Therefore, we get by (3.67):

$$J_{3,i} = \int_{I^d} c_\delta(u) \left|\log(\delta_{(i-1)}(u_{(i)}) - \delta_{(i)}(u_{(i)}))\right| du = \int_I \delta'_{(i)}(t) \left|\log(\delta_{(i-1)}(t) - \delta_{(i)}(t))\right| dt.$$

Notice that $J_{3,i} \in [0, +\infty]$ and $\sum_{i=2}^d J_{3,1} = \mathbb{J}(\delta)$. The results on $J_{1,i}$, $J_{2,i}$ and $J_{3,i}$ imply that we can decompose $H(C_\delta)$ as:

$$H(C_\delta) = \log(d!) - \sum_{i=1}^d J_{1,i} + \sum_{i=2}^d J_{2,i} - \sum_{i=2}^d J_{3,i} = \log(d!) + \sum_{i=1}^d H(\delta_{(i)}) + (d-1) - \mathbb{J}(\delta).$$

□

3.6.3 The optimization problem

Let $\delta \in \mathcal{D}^0$. Recall notation from Section 3.6.1. The problem of maximizing H over $\mathcal{C}_\delta^0$ can be written as an optimization problem (P^δ) with infinite dimensional constraints:

$$\text{maximize } H(c) \text{ subject to } \begin{cases} \mathcal{A}(c) = b^\delta, \\ c \geq 0 \text{ a.e. and } c \in L^1(I^d). \end{cases} \qquad (P^\delta)$$

Notice that if $f \in L^1(I^d)$ is non-negative and solves $\mathcal{A}(f) = b^\delta$, then f is the density of a copula. We say that a function f is feasible for (P^δ) if $f \in L^1(I^d)$, $f \geq 0$ a.e., $\mathcal{A}(f) = b^\delta$ and $H(f) > -\infty$. We say that f is an optimal solution of (P^δ) if f is feasible and $H(f) \geq H(g)$ for all g feasible. The next proposition gives conditions which ensure the existence of an optimal solution.

Proposition 3.47. *Let $\delta \in \mathcal{D}^0$. If there exists c feasible for (P^δ), then there exists a unique optimal solution to (P^δ) and it is symmetric.*

Proof. Since $\mathcal{A}(f) = b^\delta$ implies $\mathcal{A}_1(f)(1) = b_1(1)$ that is $\int_{I^d} f(x)\, dx = 1$, we can directly apply Corollary 2.3 of [25] which states that if there exists a feasible c, then there exists a unique optimal solution to (P^δ). Since the constraints of (P^δ) are symmetric, such as the functional H, we deduce that if c^* is the optimal solution, then so is c^*_π defined for $\pi \in \mathcal{S}_d$ and $u \in I^d$ as $c^*_\pi(u) = c^*(u_\pi)$. By uniqueness of the optimal solution, we deduce that $c^* = c^*_\pi$ for all permutations $\pi \in \mathcal{S}_d$; hence c^* is symmetric. □

Combining Lemmas 3.27 and 3.26 gives the following Corollary on the support of any c verifying $\mathcal{A}(c) = b^\delta$.

Corollary 3.48. *Let $\delta \in \mathcal{D}^0$. If $c \in L^1(I^d)$ is non-negative and verifies $\mathcal{A}(c) = b^\delta$, then $c = 0$ a.e. on $Z_\delta \bigcup L_\delta^c$ with L_δ defined by (3.40) and $L_\delta^c = I^d \setminus L_\delta$.*

3.6.4 Reduction of the optimization problem (P^δ)

Let $\delta \in \mathcal{D}^0$. Since the optimal solution of (P^δ) is symmetric, see Proposition 3.47, we can reduce the optimization problem by considering it on the simplex $\triangle$. We define μ to be the

Lebesgue measure restricted to $(Z_\delta^c \cap L_\delta) \cap \triangle$: $\mu(du) = \mathbf{1}_{(Z_\delta^c \cap L_\delta)\cap\triangle}(u)du$. We define, for $f \in L^1(I^d)$:

$$H^\mu(f) = -\int_{I^d} f(u)\log(f(u))\,\mu(du).$$

From Corollary 3.48 we can deduce that if $c \in L^1(I^d)$ is non-negative symmetric and solves $\mathcal{A}(c) = b^\delta$, then:

$$H(c) = d!\,H^\mu(c). \tag{3.68}$$

Let us also define, for $f \in L^1(I^d)$, $1 \le i \le d$, $r \in I$:

$$\mathcal{A}_i^\mu(c)(r) = d!\int_{I^d} c(u)\mathbf{1}_{\{u_i \le r\}}\,\mu(du).$$

We shall consider the restricted optimization problem (P_μ^δ) given by:

$$\text{maximize } H^\mu(c) \text{ subject to } \begin{cases} \mathcal{A}^\mu(c) = \delta, \\ c \ge 0 \ \mu\text{-a.e. and } c \in L^1(I^d). \end{cases} \tag{P_μ^δ}$$

We have the following equivalence between (P^δ) and (P_μ^δ). Recall that u^{OS} denotes the ordered vector of $u \in \mathbb{R}^d$.

Corollary 3.49. *Let $\delta \in \mathcal{D}^0$. If c is the optimal solution of (P^δ) then it is also an optimal solution to (P_μ^δ). If $\hat{c}$ is an optimal solution of (P_μ^δ), then c, defined by $c(u) = \hat{c}(u^{OS})\mathbf{1}_{Z_\delta^c \cap L_\delta}(u)$ is the optimal solution to (P^δ).*

Notice the Corollary implies that (P_μ^δ) has a μ-a.e. unique optimal solution: if c_1 and c_2 are two optimal solutions of (P_μ^δ) then μ-a.e. $c_1 = c_2$. Thanks to Proposition 3.47 and (3.68), Corollary 3.49 is a direct consequence of the following Lemma that establishes the connection between the constraints.

Lemma 3.50. *Let $\delta \in \mathcal{D}^0$. For $c \in L^1(I^d)$ symmetric and non-negative the following two conditions are equivalent:*

1. $\mathcal{A}(c) = b^\delta$.
2. $\mathcal{A}^\mu(c) = \delta$ *and* $c = 0$ *a.e. on* $Z_\delta \bigcup L_\delta^c$.

Proof. Assume that $\mathcal{A}(c) = b^\delta$. We have, by Corollary 3.48, that $c = 0$ a.e. on $Z_\delta \cup L_\delta^c$. This and the symmetry of c gives, for $1 \le i \le d$, $r \in I$:

$$\mathcal{A}_i^\mu(c)(r) = d!\int_{I^d} c(u)\mathbf{1}_{\{u_{(i)} \le r\}}\mathbf{1}_\triangle(u)\,du = \int_{I^d} c(u)\mathbf{1}_{\{u_{(i)} \le r\}}\,du = \delta_{(i)}(r).$$

On the other hand, let us assume that $\mathcal{A}^\mu(c) = \delta$ and $c = 0$ a.e. on $Z_\delta \cup L_\delta^c$. We have, for $1 \le i \le d$, $r \in I$:

$$\mathcal{A}_{d+i}(c)(r) = \int_{I^d} c(u)\mathbf{1}_{\{u_{(i)} \le r\}}\mathbf{1}_{Z_\delta^c \cap L_\delta}(u)\,du = d!\int_{I^d} c(u)\mathbf{1}_{\{u_i \le r\}}\,\mu(du) = \delta_{(i)}(r),$$

where we used $c = 0$ a.e. on $Z_\delta \cup L_\delta^c$ for the first equality, the symmetry of c and the definition of μ for the second, and $\mathcal{A}^\mu(c) = \delta$ for the third. Lemma 3.42 ensures then that $\mathcal{A}_i(c) = b_i$ for $1 \le i \le d$. This ends the proof. □

3.6.5 Solution for the reduced optimization problem (P_μ^δ)

Let $\delta \in \mathcal{D}^0$. We compute $(\mathcal{A}^\mu)^* : L^\infty(I)^d \to L^\infty(I^d)$ the adjoint of $\mathcal{A}^\mu$. For $\lambda = (\lambda_i, 1 \le i \le d) \in L^\infty(I)^d$ and $f \in L^1(I^d)$, we have:

$$\langle (\mathcal{A}^\mu)^*(\lambda), f\rangle = \langle \lambda, \mathcal{A}^\mu(f)\rangle = \sum_{i=1}^d \int_I \lambda_i(r)\int_{I^d} f(u)\mathbf{1}_{\{u_i \le r\}}d\mu(u)\,dr = \int_{I^d} f(u)\sum_{i=1}^d \Lambda_i(u_i)\,d\mu(u),$$

where we used the definition of the adjoint operator for the first equality, Fubini's theorem for the second, and the following definition of the functions $(\Lambda_i, 1 \le i \le d)$ for the third:

$$\Lambda_i(t) = \int_I \lambda_i(r)\mathbf{1}_{\{r \ge t\}}\, dr, \quad t \in I.$$

Thus, we have for $\lambda \in L^\infty(I)^d$ and $u = (u_1, \dots, u_d) \in I^d$:

$$(\mathcal{A}^\mu)^*(\lambda)(u) = \sum_{i=1}^d \Lambda_i(u_i). \tag{3.69}$$

We will use Theorem 2.9. from [25] on abstract entropy minimization, which we recall here, adapted to the context of (P_μ^δ).

Theorem 3.51 (Borwein, Lewis and Nussbaum)**.** *Suppose there exists $c > 0$ μ-a.e. which is feasible for (P_μ^δ). Then there exists a μ-a.e. unique optimal solution, c^*, of (P_μ^δ). Furthermore, we have $c^* > 0$ μ-a.e. and there exists a sequence $(\lambda^n, n \in \mathbb{N}^*)$ of elements of $L^\infty(I)^d$ such that:*

$$\int_{I^d} c^*(u)\, |(\mathcal{A}^\mu)^*(\lambda^n)(u) - \log(c^*(u))|\ \mu(du) \xrightarrow[n\to\infty]{} 0. \tag{3.70}$$

Now we are ready to prove that the optimal solution c^* of (P_μ^δ) is the product of measurable univariate functions.

Lemma 3.52. *Let $\delta \in \mathcal{D}^0$. Suppose that there exists $c > 0$ μ-a.e.which is feasible for (P_μ^δ). Then there exist non-negative, measurable functions $(a_i^*,\ 1 \le i \le d)$ defined on I such that $a_i^*(s) = 0$ if $\delta'_{(i)}(s) = 0$ and the function c^* defined a.e. on I^d by:*

$$c^*(u) = \frac{1}{d!}\, \mathbf{1}_{L_\delta}(u) \prod_{i=1}^d a_i^*(u_i)$$

is the optimal solution to (P_μ^δ).

Proof. According to Theorem 3.51, there exists a sequence $(\lambda^n, n \in \mathbb{N})$ of elements of $L^\infty(I)^d$ such that the optimal solution, say c^*, satisfies (3.70). This implies, thanks to (3.69), that there exist d sequences $(\Lambda_i^n, n \in \mathbb{N}^*, 1 \le i \le d)$ of elements of $L^\infty(I)$ such that the following convergence holds in $L^1(I^d, c^*\mu)$:

$$\sum_{i=1}^d \Lambda_i^n(u_i) \xrightarrow[n\to\infty]{} \log(c^*(u)). \tag{3.71}$$

We first assume that there exist Λ_i, $1 \le i \le d$ measurable functions defined on I such that μ-a.e. on S:

$$\sum_{i=1}^d \Lambda_i(u_i) = \log(c^*(u)). \tag{3.72}$$

Set $a_i^* = \sqrt[d]{d!}\exp(\Lambda_i)$ so that μ-a.e. on S:

$$c^*(u) = \frac{1}{d!} \prod_{i=1}^d a_i^*(u_i).$$

Recall $\mu(du) = \mathbf{1}_{(Z_\delta^c \cap L_\delta) \cap \triangle}(u)\, du$. From the definition (3.38) of Z_δ, we deduce that without loss of generality, we can assume that $a_i^*(u_i) = 0$ if $\delta'_{(i)}(u_i) = 0$. Therefore we obtain $c^*(u) = (1/d!)\mathbf{1}_{L_\delta}(u) \prod_{i=1}^d a_i^*(u_i)$ for $u \in I^d$.

To complete the proof, we now show that (3.72) holds for Λ_i, $1 \le i \le d$ measurable functions. We introduce the notation $u_{(-i)} = (u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_d) \in I^{d-1}$. Let us define the probability measure $P(du) = c^*(u)\mu(du)/\int_{I^d} c^*(y)\mu(dy)$ on I^d. We fix j, $1 \le j \le d$.

In order to apply Proposition 2 of [153], which ensures the existence of the limiting measurable functions Λ_i, $1 \le i \le d$, we first check that P is absolutely continuous with respect to $P_1^j \otimes P_2^j$, where $P_1^j(du_{(-j)}) = \int_{u_j \in I} P(du_{(-j)}du_j)$ and $P_2^j(du_j) = \int_{u_{(-j)} \in I^{d-1}} P(du_{(-j)}du_j)$ are the marginals of P. Notice that there exists a non-negative density function h such that $P(du) = h(u_{(-j)}, u_j)du_{(-j)}du_j$. Let $h_1(u_{(-j)}) = \int h(u_{(-j)}, u_j)du_j$ and $h_2(u_j) = \int h(u_{(-j)}, u_j)du_{(-j)}$ denote the density of the marginals P_1^j and P_2^j. Then the density of the product measure $P_1^j \otimes P_2^j$ is given by $P_1^j \otimes P_2^j(du) = h_1(u_{(-j)})h_2(u_j)du_{(-j)}du_j$. The support of the density h is noted by $T_0 = \{u \in I^d; h(u) > 0\}$, and the support of the marginals are noted by $T_1 = \{v \in I^{d-1}; h_1(v) > 0\}$ and $T_2 = \{t \in I; h_2(t) > 0\}$. With this notation, we have that a.e. $T_0 \subset T_1 \times T_2$ (that is $T_0 \cap (T_1 \times T_2)^c$ is of zero Lebesgue measure). If $A \subset I^d$ is such that $\int \mathbf{1}_A(u)h_1(u_{(-j)})h_2(u_j)du_{(-j)}du_j = 0$, then we also have $\int \mathbf{1}_{A\cap(T_1\times T_2)}(u)h_1(u_{(-j)})h_2(u_j)du_{(-j)}du_j = 0$. Since h_1h_2 is positive on $T_1 \times T_2$, this implies that $A \cap (T_1 \times T_2)$ has zero Lebesgue measure. Therefore we have:

$$\int \mathbf{1}_A(u)h(u)du = \int \mathbf{1}_{A\cap(T_1\times T_2)}(u)h(u)du + \int \mathbf{1}_{A\setminus(T_1\times T_2)}(u)h(u)du = 0,$$

since $h = 0$ a.e. on $A \setminus (T_1 \times T_2)$. This proves that P is absolutely continuous with respect to $P_1^j \otimes P_2^j$. Then according to Proposition 2 of [153], (3.71) implies that there exist measurable functions Φ_j and $\tilde{\Lambda}_j$ defined respectively on I^{d-1} and I, such that $c^*\mu$-a.e. on $\triangle$:

$$\log(c^*(u)) = \Phi_j(u_{(-j)}) + \tilde{\Lambda}_j(u_j).$$

As μ-a.e. $c^* > 0$, this equality holds μ-a.e. on S. Since we have such a representation for every $1 \le j \le d$, we can easily verify that $\log(c^*(u)) = \sum_{i=1}^d \Lambda_i(u_i)$ μ-a.e. with $\tilde{\Lambda}_j = \Lambda_j$ up to an additive constant.

□

3.6.6 Proof of Proposition 3.31

Let $\delta \in \mathcal{D}^0$. Recall that u^{OS} denotes the ordered vector of $u \in \mathbb{R}^d$. Let c be the density of a symmetric copula in $\mathbb{R}^d$ such that $\mathcal{A}(c) = b^\delta$ and c is of product form, that is, thanks to Corollary 3.49, $c(u) = c^*(u^{OS})$ with

$$c^*(u) = \frac{1}{d!}\,\mathbf{1}_{L_\delta}(u)\prod_{i=1}^d a_i^*(u_i), \quad u = (u_1, \dots, u_d) \in \triangle,$$

where a_i^*, $1 \le i \le d$ are measurable non-negative functions defined on I. In this section, we shall prove that c equals c_δ defined by (3.44); that is, for all $1 \le i \le d$, a_i^* is a.e. equal, up to a multiplicative constant, to a_i defined in (3.45). This will prove Proposition 3.31.

Recall the definitions of $g_i^{(j)}, m_i^{(j)}, d_i^{(j)}$ from Section 3.4, for $1 \le i \le d+1$. We deduce from (3.43) that:

$$c^*(u) = \frac{1}{d!}\,\mathbf{1}_{L_\delta}(u)\prod_{i=1}^d a_i^*(u_i)\mathbf{1}_{\Psi_i^\delta \cap \Psi_{i+1}^\delta}(u_i), \quad u = (u_1, \dots, u_d) \in \triangle.$$

We deduce also from Lemma 3.50 that $\mathcal{A}^\mu(c^*) = \delta$. We introduce the following family of functions:

$$B_{d+1}^*(t) = E_0^*(t) = 1,$$

and for $1 \le i \le d$, $t \in (g_i^{(j)}, d_i^{(j)})$ and $t' \in (g_{i+1}^{(j)}, d_{i+1}^{(j)})$:

$$B_i^*(t) = \int_t^{d_i^{(j)}} a_i^*(s)B_{i+1}^*(s)\,ds, \quad E_i^*(t') = \int_{g_{i+1}^{(j)}}^{t'} a_i^*(s)E_{i-1}^*(s)\,ds.$$

Recall the functions B_i, for $1 \le i \le d+1$, and E_i, for $0 \le i \le d$ defined by (3.61) and (3.63). We will prove by (downward) induction on $i \in \{1, \dots, d+1\}$ that:

$$B_i^*(t) = B_i^*(m_i^{(j)})B_i(t), \quad t \in (g_i^{(j)}, d_i^{(j)}). \tag{3.73}$$

For $i = d+1$, it trivially holds. Let us assume that (3.73) holds for $i+1$, $d \geq i \geq 1$. Recall the convention $K_{d+1} = 0$, $\delta_{(d+1)} = 0$ and $\delta_{(0)} = 1$. Arguing as in the proof of Lemma 3.45, we deduce from $\mathcal{A}_i^{\mu}(c^*) = \delta_{(i)}$ that for $r \in \Psi_i^{\delta}$:

$$\begin{aligned}
\delta_{(i)}(r) &= d! \int_{I^d} c^*(u) \mathbf{1}_{\{u_{(i)} \leq r\}} \, \mu(du) \\
&= d! \int_{I^d} c^*(u) \mathbf{1}_{\{u_{(i+1)} \leq r\}} \, \mu(du) + \int_{\triangle} \mathbf{1}_{L_\delta}(u) \prod_{j=1}^{d} \left(a_j^*(u_j) \mathbf{1}_{\Psi_j^\delta \cap \Psi_{j+1}^\delta}(u_j) \right) \mathbf{1}_{\{u_i \leq r \leq u_{i+1}\}} \, du \\
&= \delta_{(i+1)}(r) + B_{i+1}^*(r) E_i^*(r).
\end{aligned}$$

This gives on Ψ_i^{δ}:

$$\delta_{(i)} - \delta_{(i+1)} = B_{i+1}^* E_i^*. \tag{3.74}$$

Notice (3.74) holds for $i = d$ thanks to the conventions. We get on Ψ_i^{δ}:

$$\delta'_{(i)} - \delta'_{(i+1)} = -K'_{i+1} B_{i+1}^* E_i^* + B_{i+1}^* a_i^* E_{i-1}^* = -\delta'_{(i+1)} + B_{i+1}^* a_i^* E_{i-1}^*.$$

where we took the derivative in (3.74), twice the induction hypothesis for B_{i+1}^* and (3.62) for the first equality; then (3.46) and (3.74) for the second. We deduce that on Ψ_i^{δ}:

$$\delta'_{(i)} = B_{i+1}^* a_i^* E_{i-1}^*. \tag{3.75}$$

On Ψ_i^{δ}, we can divide (3.75) by (3.74) and get, thanks to (3.46):

$$\frac{a_i^* B_{i+1}^*}{B_i^*} = \frac{\delta'_{(i)}}{\delta_{(i-1)} - \delta_{(i)}} = K'_i.$$

Notice that $(B_i^*)' = -a_i^* B_{i+1}^*$. So using the representation (3.62) of B_i, we get that (3.73) holds for i. Thus (3.73) holds for $1 \leq i \leq d+1$. Then use (3.73) as well as $(B_i^*)' = -a_i^* B_{i+1}^*$ and $B_i' = -a_i B_{i+1}$ to get that for $t \in (g_i^{(j)}, d_i^{(j)}) \cap (g_{i+1}^{(k)}, d_{i+1}^{(k)})$:

$$a_i^*(t) = \frac{B_i^*(m_i^{(j)})}{B_{i+1}^*(m_{i+1}^{(k)})} a_i(t).$$

Therefore if $u = (u_1, \ldots, u_d) \in L^{\delta}$, we have:

$$\prod_{i=1}^{d} a_i^*(u_{(i)}) = \frac{B_1^*(m_1)}{B_{d+1}^*(m_{d+1})} \prod_{i=1}^{d} a_i(u_{(i)}),$$

since when $u \in L_\delta$, $u_{(i-1)}$ and $u_{(i)}$ belong to the same interval $(g_i^{(j)}, d_i^{(j)})$ for $2 \leq i \leq d$. This ensures that c_δ and c^* are densities of probability function which differ by a multiplicative constant, therefore they are equal. This ends the proof of Proposition 3.31.

3.6.7 Proof of case (a) for Theorems 3.32 and 3.37

We first consider the case $d = 2$. Let $\delta \in \mathcal{D}^0$ with $\mathbb{J}(\delta) = +\infty$. Recall $\mathbb{J}(\delta)$ is defined by (3.15). We have:

$$\begin{aligned}
\mathbb{J}(\delta) &= -\int_I \delta'_{(2)}(t) \log(2(t - \delta_{(2)}(t))) \, dt \\
&= -\log(2) - \int_I \log(t - \delta_{(2)}(t)) \, dt + \int_I (1 - \delta'_{(2)}(t)) \log(t - \delta_{(2)}(t)) \, dt \\
&= -\log(2) - \int_I \log(t - \delta_{(2)}(t)) \, dt + \Big[(t - \delta_{(2)}(t)) \log(t - \delta_{(2)}(t)) - (t - \delta_{(2)}(t)) \Big]_0^1 \\
&= -\log(2) - \int_I \log(t - \delta_{(2)}(t)) \, dt,
\end{aligned}$$

where we used $\delta_{(1)}+\delta_{(2)}=2t$ for the first equality, $\delta_{(2)}(1)=1$ and $\delta_{(2)}(0)=0$ for the second and last. In particular, we obtain that $\mathbb{J}(\delta)$ is equal to $-\log(2)+\mathcal{J}(\delta_{(2)})$, with $\mathcal{J}$ as also defined by (1) in [33]. Therefore we deduce case (a) of Theorem 3.32 (for $d=2$) from case (a) of Theorem 2.4 in [33]. Then, we get from (3.14) and Theorem 3.32 case (a) that $H_h(F)=-\infty$ for all $F\in\mathcal{L}_2^{OS}(\mathbf{F})$. This proves case (a) for Theorem 3.37 (for $d=2$).

We then consider the case $d\geq 2$. Let $\delta\in\mathcal{D}^0$ with $\mathbb{J}(\delta)=+\infty$. This implies that there exists $2\leq i\leq d$ such that $\int_I \delta'_{(i)}(t)\left|\log(\delta_{(i-1)}(t)-\delta_{(i)}(t))\right|\,dt=+\infty$. Set $\mathbf{F}=(\delta_{(i-1)},\delta_{(i)})$ and notice that $\mathbf{F}$ belongs to $\mathcal{F}_2$ as $\delta_{(i)}$ is d-Lipschitz. Since $\int_I \delta'_{(i)}(t)\left|\log(\delta_{(i-1)}(t)-\delta_{(i)}(t))\right|\,dt=+\infty$, we deduce from the first part of this Section that $\max_{F\in\mathcal{L}_2^{OS}(\mathbf{F})} H_h(F)=-\infty$.

Consider a copula C belonging to $\mathcal{C}_\delta\bigcap\mathcal{C}^{sym}$ and U a random vector on I^d with cdf C. According to Lemma 3.23 and Lemma 3.5, as C is symmetric, we have:

$$H(U^{OS})=H(U)-\log(d!)=H(C)-\log(d!).$$

It is easy to check that if $X=(X_1,\ldots,X_d)$ is a random vector on I^d and $2\leq i\leq d$, then we have $H((X_{i-1},X_i))\geq H(X)$. This implies that, for $V=(U_{i-1}^{OS},U_i^{OS})$,

$$H(V)\geq H(C)-\log(d!).$$

Since the cdf of U_ℓ^{OS} is $\delta_{(\ell)}$ as $C\in\mathcal{C}_\delta$, we deduce the cdf of V belongs to $\mathcal{L}_2^{OS}(\mathbf{F})$, and thus $H(V)=-\infty$. This implies that $H(C)=-\infty$. Thanks to Proposition 3.47 which states that the entropy is maximal on symmetric copulas, we deduce that:

$$\max_{C\in\mathcal{C}_\delta} H(C)=\max_{C\in\mathcal{C}_\delta\bigcap\mathcal{C}^{sym}} H(C)=-\infty.$$

This proves cases (a) for Theorem 3.32. Then, we get from (3.14) that $H_h(F)=-\infty$ for all $F\in\mathcal{L}_d^{OS}(\mathbf{F})$. This proves case (a) for Theorem 3.37.

3.6.8 Proof of Theorem 3.32, case (b)

Let $\delta\in\mathcal{D}$ with $\mathbb{J}(\delta)<+\infty$. Thanks to Lemma 3.18, $\mathbb{J}(\delta)<+\infty$ implies that $\delta\in\mathcal{D}^0$. By construction, c_δ introduced in Proposition 3.30 verifies $\mu-a.e.\ c_\delta>0$. The density c_δ is a feasible solution to the problem (P_μ^δ). Theorem 3.51 ensures the existence of a unique optimal solution c^*. Furthermore, by Lemma 3.52, we have that there exist non-negative, measurable functions a_i^*, $1\leq i\leq d$, such that $c^*(u)=(1/d!)\mathbf{1}_{L_\delta}(u)\prod_{i=1}^d a_i^*(u_i)$ μ-a.e. By Corollary 3.49, the optimal solution c of (P^δ) is given by, for $u=(u_1,\ldots,u_d)$:

$$c(u)=c^*(u^{OS})\mathbf{1}_{Z_\delta^c\cap L_\delta}(u)=\frac{1}{d!}\,\mathbf{1}_{L_\delta}(u)\prod_{i=1}^d a_i^*(u_{(i)})\mathbf{1}_{\{\delta'_{(i)}(u_{(i)})\neq 0\}}.$$

Since c is of product form, Proposition 3.31 yields that $c=c_\delta$ a.e., therefore C_δ is the unique copula achieving $H(C_\delta)=\max_{C\in\mathcal{C}_\delta} H(C)$.

3.7 Overview of the notations

- $\mathcal{F}_d$: set of continuous one-dimensional marginals cdf $\mathbf{F}=(\mathbf{F}_1,\ldots,\mathbf{F}_d)$ of d-dimensional order statistics, see (3.10).
- $\mathcal{F}_d^0$: set of continuous one-dimensional marginals cdf $\mathbf{F}=(\mathbf{F}_1,\ldots,\mathbf{F}_d)$ of d-dimensional abs. cont. order statistics, see Definition 3.19.
- $H_h(F)$: the relative entropy (with respect to the reference probability density h) of the random variable corresponding to the cdf F, see (3.13).
- $H(F)$: this equals $H_h(F)$ with $h=\mathbf{1}_{[0,1]}$ and F a cdf of a random variable taking values in $[0,1]^d$.

- $\mathbb{J}(\mathbf{F})$: the quantity appearing in the expression of the entropy of the maximum entropy distribution of order statistics with marginals cdf $\mathbf{F} \in \mathcal{F}_d$, see (3.15).
- $\mathcal{L}_d$: set of cdf's on $\mathbb{R}^d$ with continuous one-dimensional marginals cdf.
- $\mathcal{L}_d^0$: set of absolutely continuous cdf's on $\mathbb{R}^d$.
- $\mathcal{L}_d^{OS}$: set of cdf's of d-dimensional order statistics with continuous one dimensional marginals cdf.
- $\mathcal{L}_d^{OS}(\mathbf{F})$: set of cdf's of d-dimensional order statistics with marginals cdf $\mathbf{F}$, see (3.11).
- $\mathcal{L}_d^{sym}$: set of symmetric cdf's on $\mathbb{R}^d$ with continuous one-dimensional marginals cdf.
- F^{sym}: the symmetrization of the cdf F, see (3.16).
- $S_{\mathbf{F}}$: symmetrizing operator on copulas, associated to the marginals cdf $\mathbf{F}$, see Definition 3.3.
- $\mathcal{C}$: set of all copulas.
- $\mathcal{C}^0$: set of absolutely continuous copulas.
- $\mathcal{C}^{OS}(\mathbf{F})$: set of copulas of order statistics with marginals cdf $\mathbf{F}$, see (3.12).
- $\mathcal{C}^{sym}$: set of symmetric (permutation invariant) copulas.
- $\mathcal{C}^{sym}(\mathbf{F})$: image of the set $\mathcal{C}^{OS}(\mathbf{F})$ by the operator $S_{\mathbf{F}}$, see (3.19). It is the set of symmetric copulas with multidiagonal $\delta^{\mathbf{F}}$.
- $\mathcal{C}_\delta$: set of copulas with multidiagonal δ, see Section 3.3.2.
- $\mathcal{C}_\delta^0$: set of abs. cont. copulas with multidiagonal δ, see Section 3.3.2.
- $\mathcal{D}$: set of multidiagonals of copulas, see Section 3.3.2.
- $\mathcal{D}^0$: set of multidiagonals of abs. cont. copulas, see Section 3.3.2.
- $\Psi_i^{\mathbf{F}}$: set of points $t \in \mathbb{R}$ for which the marginals cdf $\mathbf{F} = (\mathbf{F}_1, \dots, \mathbf{F}_d)$ verify $\mathbf{F}_{i-1}(t) > \mathbf{F}_{\mathbf{i}}(t)$, see (3.29).
- $T^{\mathbf{F}}$: set of points $u = (u_1, \dots, u_d) \in I^d$ for which $\mathbf{F}_1^{-1}(u_1) \leq \dots \leq \mathbf{F}_d^{-1}(u_d)$, see (3.33). The density of all copulas in $\mathcal{C}^{OS}(\mathbf{F})$ vanishes outside $T^{\mathbf{F}}$.
- $L^{\mathbf{F}}$: set of ordered vectors $x \in \mathbb{R}^d$ such that the marginals cdf's $\mathbf{F} = (\mathbf{F}_1, \dots, \mathbf{F}_d)$ verify $\mathbf{F}_{i-1}(t) > \mathbf{F}_i(t)$ for all $t \in (x_{i-1}, x_i)$, $2 \leq i \leq d$, see (3.51). The density of any abs. cont. cdf in $\mathcal{L}_d^{OS}(\mathbf{F})$ vanishes outside $L^{\mathbf{F}}$.
- L_δ: set of points $u = (u_1, \dots, u_d) \in I^d$ for which all points $t \in (u_{(i-1)}, u_{(i)})$ verify $\delta_{(i-1)}(t) > \delta_{(i)}(t)$ for all $2 \leq i \leq d$, see (3.40). The density of any copula in $\mathcal{C}_\delta^0$ vanishes outside L_δ.
- Z_δ: set of points $u = (u_1, \dots, u_d) \in I^d$ such that $\delta'_{(i)}(u_{(i)}) = 0$ for some $1 \leq i \leq d$, see (3.38). The density of any copula in $\mathcal{C}_\delta^0$ vanishes on Z_δ.

Chapitre 4

Application pour la quantification d'incertitude

4.1 Introduction

L'étude modélise une plaque d'acier sur laquelle on fait une ligne de soudure. Un calcul physique correspond à l'enchaînement de trois calculs :

- un calcul thermique qui définit l'historique de température,
- un calcul métallurgique (éventuellement, qui définit la transformation subie par l'acier due à l'historique de température précédent),
- un calcul mécanique qui estime la carte des contraintes résiduelles dans l'acier suite à la soudure.

Une analyse de sensibilité effectuée sur les paramètres mécaniques permet de déterminer les paramètres sur lesquels l'analyste doit porter son attention en priorité. Les paramètres mécaniques considérés sont les suivants :

- E : module de Young de l'acier,
- α : coefficient de dilatation thermique,
- σ_m : limite d'élasticité,
- D_{sde} : pente d'écrouissage.

Chacun des paramètres $(E, \alpha, \sigma_m, D_{sde})$ est une fonction monotone de la température T : croissante pour α, décroissante pour (E, σ_m, D_{sde}). La plage de températures considérée est $[20°C, 1200°C]$, discrétisée en un nombre n de températures ($n \sim 10$).

4.2 Modélisation actuelle

La variation des paramètres est modélisée comme suit. Pour chaque paramètre P et température T , on considère :

- la valeur nominale de P à la température T (donnée expérimentale), notée par $P_0(T)$;
- une amplitude maximale $Err^P(T)$ de P autour de $P_0(T)$, qui dépend de manière polynomiale de T :
$$Err^P(T) = \lambda^P (T - T_{min})^k + Err^P_{min} \tag{4.1}$$
où $(\lambda^P, T_{min}, Err^P_{min})$ sont des paramètres fixés, $k = 1$ ou $k = 2$;
- la variation de P autour de sa valeur nominale est donnée par :
$$P(T) = P_0(T) \left(1 + A^P \times Err^P(T)\right) \tag{4.2}$$
où A^P est une variable aléatoire uniforme sur $[-1, 1]$. A^P ne dépend pas de T.

Ainsi, la valeur de la variable aléatoire A^P permet de déterminer la courbe $P(T)$ qui respecte bien la monotonie imposée par la physique (quand les paramÃ¨tres sont correctement choisis). Les variables aléatoires $(A^E, A^\alpha, A^{\sigma_m}, A^{D_{sde}})$ sont supposées indépendantes. L'analyse de sensibilité consiste donc à étudier la sensibilité des résultats de sortie (carte des contraintes résiduelles) à

ces quatre paramètres. L'avantage de cette méthode est de permettre la réduction du nombre de variables dans l'analyse de sensibilité : la sensibilité des contraintes résiduelles est calculée par rapport aux quatre variables $(A^E, A^\alpha, A^{\sigma_m}, A^{D_{sde}})$ et non plus aux $4n$ variables initiales. L'inconvénient de cette méthode est qu'elle impose une relation de monotonie particulière en fonction de la température pour chaque paramètre, donnée par la relation (4.2), une fois l'aléa $(A^E, A^\alpha, A^{\sigma_m}, A^{D_{sde}})$ réalisé. De plus, pour les paramètres décroissants nous devons nous assurer que la fonction $P_0(T)(1 + Err^P(T))$ est bien décroissante, car cette propriété n'est pas automatiquement assurée par le modèle.

4.3 Données de la littérature sur les paramètres mécaniques

Le nombre de bases de donnée sur les paramètres mécaniques du soudage est limité. On fait référence à la thèse de Depradeux [62] qui rassemble les différents sources de données sur ces paramètres. Figure 4.1 illustre par exemple la variation des profils de la limite d'élasticité issus de différentes bases de donnée. La modélisation actuelle utilise également les valeurs de paramètres obtenues dans ce document comme valeurs nominales.

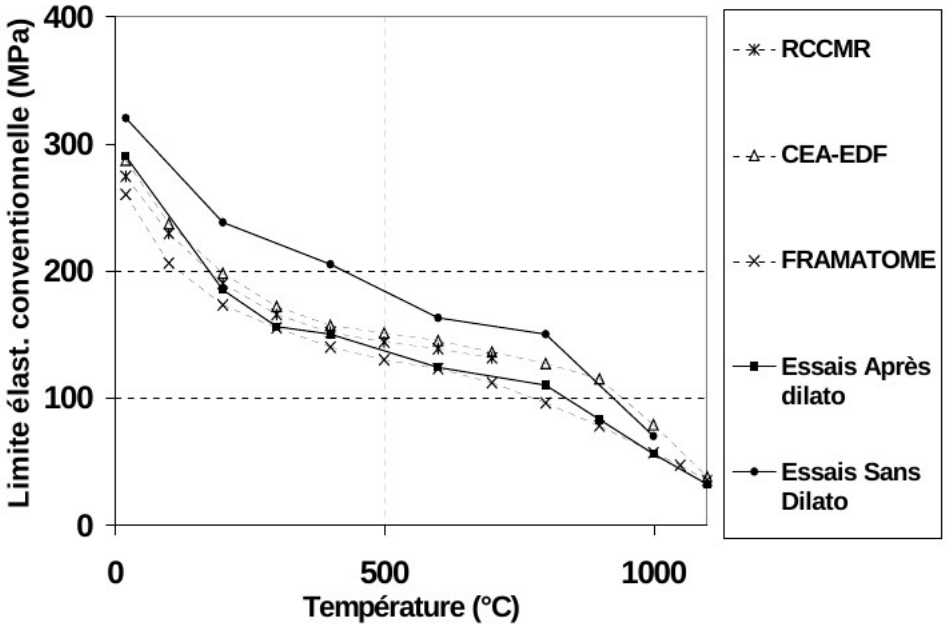


FIGURE 4.1 – Comparaison des limites conventionnelles d'élasticité σ_m avec des valeurs issues de différentes bases de donnée existantes, figure de [62]

On observe qu'il y a une forte variation à chaque instance de température, et que les courbes ont différentes formes qui peuvent éventuellement se croiser. Cette caractéristique des paramètres n'est pas prise en compte par la modélisation actuelle qui impose la forme de la courbe simulée, donc elle impose également que la différence entre la valeur nominale et la valeur modélisée a la même signe pour chaque température. Ces problèmes font appel à une modélisation différente qui peut assurer une plus grande variation des profils de paramètres toujours conservant la monotonie et assurant un contrôle sur les marginales à chaque température.

4.4 Modélisation proposée à l'aide d'une copule d'entropie maximale

Afin de relâcher la contrainte sur la forme de la courbe des paramètres, nous proposons une approche différente pour modéliser chacun des $4n$ paramètres $(E(T_i), \alpha(T_i), \sigma_m(T_i), D_{sde}(T_i))_{i=1}^n$ qui conserve les marginales uniformes des $P(T_i)$ autour de sa valeur nominale $P_0(T_i)$ et qui respecte la contrainte de la monotonie sans rajouter aucune autre. La distribution de la variable aléatoire $P(T_i)$ est uniforme sur $I^{T_i} = [P^{min}(T_i), P^{max}(T_i)]$ avec :

$$P^{min}(T_i) = P_0(T_i)(1 - Err^P(T_i)) \quad \text{et} \quad P^{max}(T_i) = P_0(T_i)(1 + Err^P(T_i))$$

La longueur de l'intervalle I^{T_i} sera notée par L_i. Notons par F_i la fonction de répartition de $P(T_i)$ donnée par, pour $x \in \mathbb{R}$:

$$F_i(x) = \frac{1}{L_i}\left(x - P^{min}(T_i)\right)\mathbf{1}_{I^{T_i}}(x) + \mathbf{1}_{(P^{max}(T_i),+\infty)}(x).$$

Pour pouvoir utiliser les résultats sur la copule d'entropie maximale des statistiques d'ordre, on doit s'assurer que les conditions suivantes soient vérifiées par les marginales.

$$F_{i-1}(t) > F_i(t) \tag{4.3}$$

pour tout $t \in (a_{i-1}, b_i)$, avec $a_{i-1} = \sup\{x \in \mathbb{R}; F_{(i-1)}(x) = 0\}$ et $b_i = \inf\{x \in \mathbb{R}; F_{(i)}(x) = 1\}$; d'autre part une condition d'intégrabilité donnée par :

$$\sum_{i=2}^{d} \int_{\mathbb{R}} f_i(t)\,|\log(F_{i-1}(t) - F_i(t))|\,dt < +\infty. \tag{4.4}$$

Pour assurer (4.3), il suffit que les quantités $P^{min}(T_i), P^{max}(T_i)$ vérifient pour $i = 2, \ldots, n$:

$$P^{min}(T_{i-1}) \leq P^{min}(T_i) \quad \text{et} \quad P^{max}(T_{i-1}) \leq P^{max}(T_i). \tag{4.5}$$

La condition (4.4) est vérifiée une fois qu'on a (4.5). Alors la densité f_P^* de la loi jointe d'entropie maximale pour $P = (P(T_1), \ldots, P(T_n))$ s'écrit d'après (3.54), pour $x_1, \ldots, x_n \in \mathbb{R}$:

$$f_P^*(x_1, \ldots, x_n) = \begin{cases} \frac{1}{L_1} \prod_{i=2}^n \frac{1}{L_i(F_{i-1}(x_i) - F_i(x_i))}\, \mathrm{e}^{-\int_{x_{i-1}}^{x_i} \frac{\mathbf{1}_{I^{T_i}}(s)}{L_i(F_{i-1}(s) - F_i(s))}\, ds} & \text{si } \forall i : x_i \in I^{T_i}, \\ 0 & \text{sinon.} \end{cases} \tag{4.6}$$

4.5 Simulation de P

Une réalisation de P peut être engendrée de manière séquentielle :

1. On engendre U_1 uniforme sur $[0, 1]$, et on met $P(T_1) = L_1 U_1 + P^{min}(T_1)$.
2. On engendre $P(T_i)$, $i = 2, \ldots, n$ à l'aide des U_i uniformes sur $[0, 1]$ de la façon suivante :
   (a) si $P^{max}(T_{i-1}) < P^{min}(T_i)$, alors $P(T_i) = L_i U_i + P^{min}(T_i)$;
   (b) si $P^{min}(T_i) \leq P^{max}(T_{i-1})$ on utilise la fonction de répartition conditionnelle de $P(T_i)$ étant donnée la valeur de $P(T_{i-1})$:

$$F_i(x_i | P(T_{i-1}) = x_{i-1}) = 1 - \mathrm{e}^{-\int_{x_{i-1}}^{x_i} \frac{\mathbf{1}_{I^{T_i}}(s)}{L_i(F_{i-1}(s) - F_i(s))}\, ds} \qquad \text{pour } x_{i-1} \leq x_i \leq P^{max}(T_i),$$

   et on met $P(T_i) = F_i^{-1}(U_i | P(T_{i-1}))$.

Quand on souhaite donner la forme analytique de la fonction $F_i(.|P(T_{i-1}))$ requise en (2b), il faut distinguer entre deux cas selon la valeur de $P(T_{i-1})$. On définit la constante $c_i = \frac{L_i}{L_{i-1}} - 1$. Si $P(T_{i-1}) = x_{i-1} < P^{min}(T_i)$ et $c_i \neq 0$, on obtient, pour $x_i \in I^{T_i}$:

$$F_i(x_i|P(T_{i-1}) = x_{i-1}) = \begin{cases} 1 - \left(\frac{c_i P^{min}(T_i) - d_i}{c_i x_i - d_i}\right)^{\frac{1}{c_i}} & \text{si } x_i \in [P^{min}(T_i), P^{max}(T_{i-1})], \\ 1 - g_i \frac{P^{max}(T_i) - x_i}{P^{max}(T_i) - P^{max}(T_{i-1})} & \text{si } x_i \in [P^{max}(T_{i-1}), P^{max}(T_i)], \end{cases}$$

avec des constantes $d_i = \frac{L_i}{L_{i-1}} P^{min}(T_{i-1}) - P^{min}(T_i)$ et $g_i = \left(\frac{c_i P^{min}(T_i) - d_i}{c_i P^{max}(T_{i-1}) - d_i}\right)^{\frac{1}{c_i}}$. Si $c_i = 0$ alors l'expression prend la forme :

$$F_i(x_i|P(T_{i-1}) = x_{i-1}) = \begin{cases} 1 - \mathrm{e}^{(x_i - P^{min}(T_i))/d_i} & \text{si } x_i \in [P^{min}(T_i), P^{max}(T_{i-1})], \\ 1 - \hat{g}_i \frac{P^{max}(T_i) - x_i}{P^{max}(T_i) - P^{max}(T_{i-1})} & \text{si } x_i \in [P^{max}(T_{i-1}), P^{max}(T_i)], \end{cases}$$

avec $\hat{g}_i = \mathrm{e}^{(P^{max}(T_{i-1}) - P^{min}(T_i))/d_i}$. Dans le cas $x_{i-1} \geq P^{min}(T_i)$, on obtient, pour $c_i \neq 0$ et $x_i \in I^{T_i}$:

$$F_i(x_i|P(T_{i-1}) = x_{i-1}) = \begin{cases} 0 & \text{si } x_i \in [P^{min}(T_i), x_{i-1}], \\ 1 - \left(\frac{c_i x_{i-1} - d_i}{c_i x_i - d_i}\right)^{\frac{1}{c_i}} & \text{si } x_i \in [x_{i-1}, P^{max}(T_{i-1})], \\ 1 - h_i \frac{P^{max}(T_i) - x_i}{P^{max}(T_i) - P^{max}(T_{i-1})} & \text{si } x_i \in [P^{max}(T_{i-1}), P^{max}(T_i)], \end{cases}$$

avec $h_i = \left(\frac{c_i x_{i-1} - d_i}{c_i P^{max}(T_{i-1}) - d_i}\right)^{\frac{1}{c_i}}$. Quand $c_i = 0$, on a :

$$F_i(x_i|P(T_{i-1}) = x_{i-1}) = \begin{cases} 0 & \text{si } x_i \in [P^{min}(T_i), x_{i-1}], \\ 1 - \mathrm{e}^{(x_i - x_{i-1})/d_i} & \text{si } x_i \in [x_{i-1}, P^{max}(T_{i-1})], \\ 1 - \hat{h}_i \frac{P^{max}(T_i) - x_i}{P^{max}(T_i) - P^{max}(T_{i-1})} & \text{si } x_i \in [P^{max}(T_{i-1}), P^{max}(T_i)], \end{cases}$$

avec $\hat{h}_i = \mathrm{e}^{(P^{max}(T_{i-1}) - x_{i-1})/d_i}$. La fonction inverse $F_i^{-1}(.|P(T_{i-1}))$ est donnée par, quand $x_{i-1} < P^{min}(T_i)$ et $c_i \neq 0$:

$$F_i^{-1}(u_i|P(T_{i-1}) = x_{i-1}) = \begin{cases} \frac{c_i P^{min}(T_i) - d_i}{c_i (1-u_i)^{c_i}} + \frac{d_i}{c_i} & \text{si } u_i \in [0, 1 - g_i], \\ P^{max}(T_i) - \frac{(1-u_i)(P^{max}(T_i) - P^{max}(T_{i-1}))}{g_i} & \text{si } u_i \in [1 - g_i, 1]. \end{cases}$$

Si $c_i = 0$, alors l'inverse est donnée par :

$$F_i^{-1}(u_i|P(T_{i-1}) = x_{i-1}) = \begin{cases} P^{min}(T_i) + d_i \log(1 - u_i) & \text{si } u_i \in [0, 1 - \hat{g}_i], \\ P^{max}(T_i) - \frac{(1-u_i)(P^{max}(T_i) - P^{max}(T_{i-1}))}{\hat{g}_i} & \text{si } u_i \in [1 - \hat{g}_i, 1]. \end{cases}$$

Dans le cas $x_{i-1} \geq P^{min}(T_i)$ et $c_i \neq 0$, l'inverse est donnée par :

$$F_i^{-1}(u_i|P(T_{i-1}) = x_{i-1}) = \begin{cases} \frac{c_i x_{i-1} - d_i}{c_i (1-u_i)^{c_i}} + \frac{d_i}{c_i} & \text{si } u_i \in [0, 1 - h_i], \\ P^{max}(T_i) - \frac{(1-u_i)(P^{max}(T_i) - P^{max}(T_{i-1}))}{h_i} & \text{si } u_i \in [1 - h_i, 1]. \end{cases}$$

Finalement, quand $c_i = 0$, on a :

$$F_i^{-1}(u_i|P(T_{i-1}) = x_{i-1}) = \begin{cases} x_{i-1} + d_i \log(1 - u_i) & \text{si } u_i \in [0, 1 - \hat{h}_i], \\ P^{max}(T_i) - \frac{(1-u_i)(P^{max}(T_i) - P^{max}(T_{i-1}))}{\hat{h}_i} & \text{si } u_i \in [1 - \hat{h}_i, 1]. \end{cases}$$

La forme analytique des inverses nous permet de simuler le vecteur P sans complexifier les calculs. La figure 4.2 illustre la différences entre les réalisations de P en utilisant les deux approches décrites ci-dessus. L'approche proposée n'impose pas de modèle particulier à la monotonie des paramètres, ce qui entraîne donc des profils plus réalistes des paramètres mécaniques.

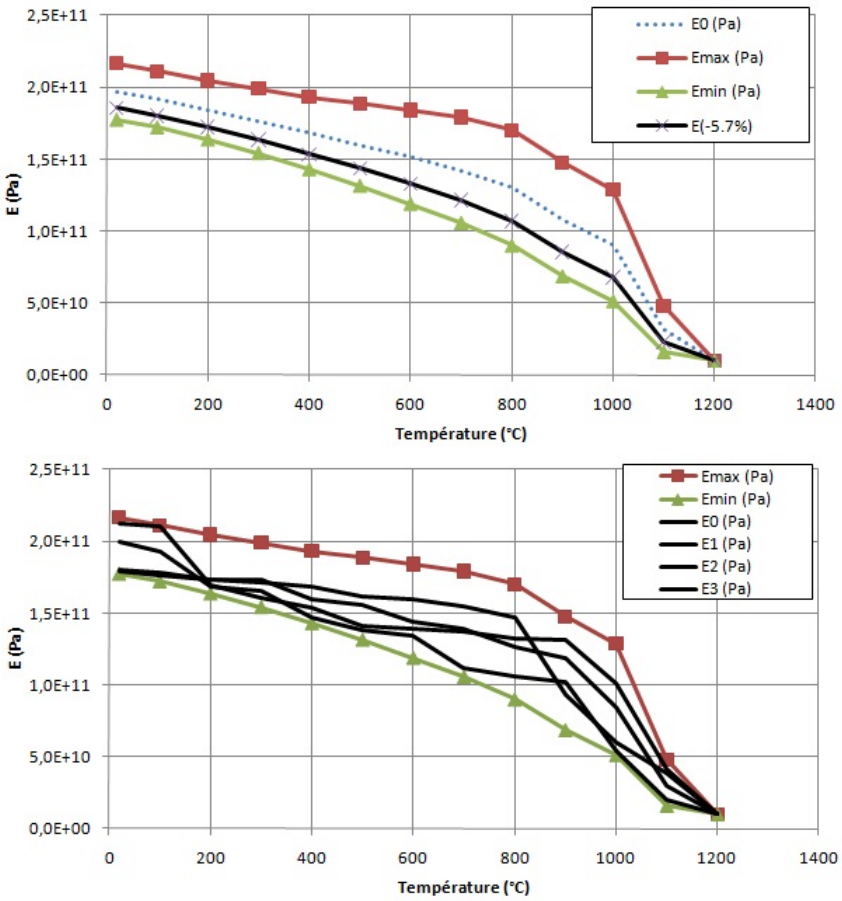


Figure 4.2 – Profils du module de Young E issues de l'approche initiale (en haut) et de l'approche innovante (en bas).

4.6 Conclusion

Dans cette communication, nous avons étudié les distributions de statistiques d'ordre sous contrainte de marginales fixées. Les marginales étant imposées, la copule du vecteur aléatoire comprend toutes les informations sur la dépendance entre ses composants, ainsi que sur la loi jointe. La présence d'une contrainte de relation d'ordre limite l'espace des copules compatibles. Nous avons donné une caractérisation des copules des statistiques d'ordre avec des marginales fixées, ce qui permet de calculer la densité de la copule optimale maximisant l'entropie (donc l'incertitude) du vecteur aléatoire. Nous avons illustré à travers un exemple de simulation des paramètres mécaniques du soudage que les résultats théoriques peuvent être bien implémentés dans un contexte industriel. Les marginales uniformes dans ce cas particulier mènent à un schéma de simulation explicite, qui n'est pas toujours assuré lorsqu'on se donne des marginales différentes. L'étude des méthodes de simulation dans le cas général fera partie des travaux futurs et l'approche proposée sera appliquée à d'autres études de quantification d'incertitude liées aux activités d'EDF.

Part II

Nonparametric estimation of maximum entropy distributions of order statistics

Chapter 5

Optimal exponential bounds for aggregation of estimators for the Kullback-Leibler loss

5.1 Introduction

The pure aggregation framework with deterministic estimators was first established in [135] for nonparametric regression with random design. Given N estimators $f_k, 1 \leq k \leq N$ and a sample $X = (X_1, \ldots, X_n)$ from the model f, the problem is to find an aggregated estimate $\hat{f}$ which performs nearly as well as the best f_μ, $\mu \in \mathcal{U}$, where:

$$f_\mu = \sum_{k=1}^{N} \mu_k f_k,$$

and $\mathcal{U}$ is a certain subset of $\mathbb{R}^N$ (we assume that linear combinations of the estimators are valid candidates). The performance of the estimator is measured by a loss function L. Common loss functions include L^p distance (with $p = 2$ in most cases), Kullback-Leibler or other divergences, Hellinger distance, etc. The aggregation problem can be formulated as follows: find an aggregate estimator $\hat{f}$ such that for some $C \geq 1$ constant, $\hat{f}$ satisfies an oracle inequality in expectation, i.e.:

$$\mathbb{E}\left[L(f, \hat{f})\right] \leq C \min_{\mu \in \mathcal{U}} L(f, f_\mu) + R_{n,N}, \tag{5.1}$$

or in deviation, i.e. for $\varepsilon > 0$ we have with probability greater than $1 - \varepsilon$:

$$L(f, \hat{f}) \leq C \min_{\mu \in \mathcal{U}} L(f, f_\mu) + R_{n,N,\varepsilon}, \tag{5.2}$$

with remainder terms $R_{n,N}$ and $R_{n,N,\varepsilon}$ which do not depend on f or $f_k, 1 \leq k \leq N$. If $C = 1$, then the oracle inequality is sharp.

Three types of problems were identified depending on the choice of $\mathcal{U}$. In the model selection problem, the estimator mimics the best estimator amongst $f_1, \ldots, f_N$, that is $\mathcal{U} = \{e_k, 1 \leq k \leq N\}$, with $e_k = (\mu_j, 1 \leq j \leq N) \in \mathbb{R}^N$ the unit vector in direction k given by $\mu_j = \mathbf{1}_{\{j=k\}}$. In the convex aggregation problem, f_μ are the convex combinations of $f_k, 1 \leq k \leq N$, i.e. $\mathcal{U} = \Lambda^+ \subset \mathbb{R}^N$ with:

$$\Lambda^+ = \{\mu = (\mu_k, 1 \leq k \leq N) \in \mathbb{R}^N, \mu_k \geq 0 \text{ and } \sum_{1 \leq k \leq N} \mu_k = 1\}. \tag{5.3}$$

Finally in the linear aggregation problem we take $\mathcal{U} = \mathbb{R}^N$, the entire linear span of the initial estimators.

Early papers usually consider the L^2 loss in expectation as in (5.1). For the regression model with random design, optimal bounds for the L^2 loss in expectation for model selection aggregation was considered in [178] and [174], for convex aggregation in [107] with improved

results for large N in [180], and for linear aggregation in [168]. These results were extended to the case of regression with fixed design for the model selection aggregation in [53] and [54], and for affine estimators in the convex aggregation problem in [52]. A unified aggregation procedure which achieves near optimal loss for all three problems simultaneously was proposed in [29].

For density estimation, early results include [41] and [179] which independently considered the model selection aggregation under the Kullback-Leibler loss in expectaion. They introduced the progressive mixture method to give a series of estimators which verify oracle inequalities with optimal remainder terms. This method was later generalized as the mirror averaging algorithm in [108] and applied to various problems. Corresponding lower bounds which ensure the optimality of this procedure was shown in [120]. The convex and linear aggregation problems for densities under the L^2 loss in expectation were considered in [150].

While a lot of papers considered the expected value of the loss, relatively few papers address the question of optimality in deviation, that is with high probability as in (5.2). For the regression problem with random design, [8] shows that the progressive mixture method is deviation sub-optimal for the model selection aggregation problem, and proposes a new algorithm which is optimal for the L^2 loss in deviation and expectation as well. Another deviation optimal method based on sample splitting and empirical risk minimization on a restricted domain was proposed in [121]. For the fixed design regression setting, [149] considers all three aggregation problems in the context of generalized linear models and gives constrained likelihood maximization methods which are optimal in both expectation and deviation with respect to the Kullback-Leibler loss. More recently, [50] extends the results of [149] for model selection by introducing the Q-aggregation method and giving a greedy algorithm which produces a sparse aggregate achieving the optimal rate in deviation for the L^2 loss. More general properties of this method applied to other aggregation problems as well are discussed in [51].

For the density estimation, optimal bounds in deviation with respect to the L^2 loss for model selection aggregation are given in [17]. The author gives a non-asymptotic sharp oracle inequality under the assumption that f and the estimators $f_k, 1 \leq k \leq N$ are bounded, and shows the optimality of the remainder term by providing the corresponding lower bounds as well. The penalized empirical risk minimization procedure introduced in [17] inspired our current work. Here, we consider a more general framework which incorporates, as a special case, the density estimation problem. Moreover, we give results in deviation for the Kullback-Leibler loss instead of the L^2 loss considered in [17].

Linear aggregation of lag window spectral density estimators with L^2 loss was studied in [43]. The method we propose is more general as it can be applied to any set of estimators f_k, $1 \leq k \leq N$, not only kernel estimators. However we consider the model selection problem, which is weaker than the linear aggregation problem. Also, this paper concerns optimal bounds in deviation for the Kullback-Leibler loss instead of the L^2 loss in expectaion.

We now present our main contributions. We propose aggregation schemes for the estimation of probability densities on $\mathbb{R}^d$ and the estimation of spectral densities of stationary Gaussian processes. We consider model selection type aggregation for the Kullback-Leibler loss in deviation. For positive, integrable functions p, q, let $D(p\|q)$ denote the generalized Kullback-Leibler divergence given by:

$$D(p\|q) = \int p \log(p/q) - \int p + \int q. \tag{5.4}$$

This is a Bregman-divergence, therefore $D(p\|q)$ is non-negative and $D(p\|q) = 0$ if and only if a.e. $p = q$. The Kullback-Leibler loss of an estimator $\hat{f}$ is given by $D(f||\hat{f})$. For initial estimators $f_k, 1 \leq k \leq N$, the aggregate estimator $\hat{f}$ verifies the following sharp oracle inequality for every f belonging to a large class of functions $\mathcal{F}$, with probability greater than $1 - \exp(-x)$ for all $x > 0$:

$$D\left(f\|\hat{f}\right) \leq \min_{1\leq k\leq N} D(f\|f_k) + R_{n,N,x}. \tag{5.5}$$

We propose two methods of convex aggregation for non-negative estimators, see Propositions 5.3 and 5.3. Contrary to the usual approach of giving an aggregate estimator which is a linear or convex combination of the initial estimators, we consider an aggregation based on a convex

combination of the logarithms of these estimators. The *convex aggregate estimators* $\hat{f} = f^D_{\hat{\lambda}}$ and $\hat{f} = f^S_{\hat{\lambda}}$ with $\hat{\lambda} = \hat{\lambda}(X_1, \ldots, X_n) \in \Lambda^+$ maximizes a penalized maximum likelihood criterion. The exact form of the convex aggregates $f^D_{\hat{\lambda}}$ and $f^S_{\hat{\lambda}}$ will be precised in later sections for each setup.

The first method concerns estimators with a given total mass and produces an aggregate $f^D_{\hat{\lambda}}$ which has also the same total mass. This method is particularly adapted for density estimation as it provides an aggregate which is also proper density function. We use this method to propose an adaptive nonparametric density estimator for maximum entropy distributions of order statistics in [37]. The second method, giving the aggregate $f^S_{\hat{\lambda}}$, does not have the mass conserving feature, but can be applied to a wider range of statistical estimation problems, in particular to spectral density estimation. We show that both procedures give an aggregate which verifies a sharp oracle inequality with a bias and a variance term. When applied to density estimation, we obtain sharp oracle inequalities with the optimal remainder term of order $\log(N)/n$, that is we have (5.5) with:

$$R_{n,N,x} = \beta \frac{\log(N) + x}{n},$$

with β depending only on the infinity norm of the logarithms of f and $f_k, 1 \leq k \leq N$, see Theorem 5.6. In the case of spectral density estimation, we need to suppose a minimum of regularity for the logarithm of the true spectral density and the estimators. We require that the logarithms of the functions belong to the periodic Sobolev space W_r with $r > 1/2$. We show that this also implies that the spectral densities itself belong to W_r. We obtain (5.5) with:

$$R_{n,N,x} = \beta \frac{\log(N) + x}{n} + \frac{\alpha}{n},$$

where β and α constants which depend only on the regularity and the Sobolev norm of the logarithms of f and $f_k, 1 \leq k \leq N$, see Theorem 5.10.

To show the optimality in deviation of the aggregation procedures, we give the corresponding tight lower bounds as well, with the same remainder terms, see Propositions 5.13 and 5.14. This complements the results of [120] and [17] obtained for the density estimation problem. In [120] the lower bound for the expected value of the Kullback-Leibler loss was shown with the same order for the remainder term, while in [17] similar results were obtained in deviation for the L^2 loss.

The rest of the paper is organised as follows. In Section 5.2 we introduce the notation and give the basic definitions used in the rest of the paper. We present the two types of convex aggregation method for the logarithms in Sections 5.3.1 and 5.3.1. For the model selection aggregation problem, we give a general sharp oracle inequality in deviation for the Kullback-Leibler loss for each method. In Section 5.3.2 we apply the methods for the probability density and the spectral density estimation problems. The results on the corresponding lower bounds can be found in Section 5.4 for both problems. We summarize the properties of Toeplitz matrices and periodic Sobolev spaces in the Appendix.

5.2 Notations

Let $\mathcal{B}_+(\mathbb{R}^d)$, $d \geq 1$, be the set of non-negative measurable real function defined on $\mathbb{R}^d$ and $h \in \mathcal{B}_+(\mathbb{R}^d)$ be a reference probability density. For $f \in \mathcal{B}_+(\mathbb{R}^d)$, we define:

$$g_f = \log(f/h), \tag{5.6}$$

with the convention that $\log(0/0) = 0$. Notice that we have $\|g_f\|_\infty < \infty$ if and only if f and h have the same support $\mathcal{H} = \{h > 0\}$. We consider the subset $\mathcal{G}$ of the set of non-negative measurable functions with support $\mathcal{H} = \{h > 0\}$:

$$\mathcal{G} = \{f \in \mathcal{B}_+(\mathbb{R}^d);\ \|g_f\|_\infty < +\infty\}.$$

For $f \in \mathcal{G}$, we set:

$$m_f = \int f, \quad \psi_f = -\int g_f\, h \quad \text{and} \quad t_f = g_f + \psi_f, \tag{5.7}$$

and we get $\int t_f\, h = 0$ as well as the inequalities:

$$m_f \leq \mathrm{e}^{\|g_f\|_\infty}, \quad |\psi_f| \leq \|\, g_f \,\|_\infty, \quad \|\, t_f \,\|_\infty \leq 2\, \|\, g_f \,\|_\infty \quad \text{and} \quad \psi_f + \log(m_f) \leq \|\, t_f \,\|_\infty . \tag{5.8}$$

Notice that the Kullback-Leibler divergence $D\,(f' \| f)$, defined in (5.4), is finite for any function $f', f \in \mathcal{G}$. When there is no confusion, we shall write g, m, ψ and t for g_f, m_f, ψ_f and t_f.

We consider a probabilistic model $\mathcal{P} = \{\mathrm{P}_f;\, f \in \mathcal{F}(L)\}$, with $\mathcal{F}(L)$ a subset of $\mathcal{G}$ with additional constraints (such as smoothness or integral condition) and P_f a probability distribution depending on f. In the sequel, the model P_f corresponds to a sample of i.i.d. random variables with density f (Section 5.3.1) or a sample from a stationary Gaussian process with spectral density f (Section 5.3.1). Suppose we have $(f_k, 1 \leq k \leq N)$, which are N distinct estimators of the function $f \in \mathcal{F}(L)$ such that there exists $K > 0$ (possibly different from L) for which $f_k \in \mathcal{F}(K)$ for $1 \leq k \leq N$, as well as a sample $X = (X_1, \ldots, X_n)$, $n \in \mathbb{N}^*$ with distribution P_f. We shall propose two convex aggregation estimator of f, based on these estimators and the available sample, that behaves, with high probability, as well as the best initial estimator f_{k^*} in terms of the Kullback-Leibler divergence, where k^* is defined as:

$$k^* = \operatorname*{argmin}_{1 \leq k \leq N} D\,(f \| f_k)\,. \tag{5.9}$$

For $1 \leq k \leq N$, we set $g_k = g_{f_k}$, $m_k = m_{f_k}$, $\psi_k = \psi_{f_k}$ and $t_k = t_{f_k}$. Notice that:

$$f = \exp(g)\, h = \exp(t - \psi)\, h \quad \text{and} \quad f_k = \exp(g_k)\, h = \exp(t_k - \psi_k)\, h. \tag{5.10}$$

We denote by I_n an integrable estimator of the function f measurable with respect to the sample $X = (X_1, \ldots, X_n)$. The estimator I_n may be a biased estimator of f. We note $\bar{f}_n$ the expected value of I_n:

$$\bar{f}_n = \mathbb{E}[I_n].$$

We fix some additional notation. For a measurable function p on $\mathbb{R}^d$ and a measure Q on $\mathbb{R}^d$ (resp. a measurable function q on $\mathbb{R}^d$), we write $\langle p, Q\rangle = \int p(x) Q(dx)$ (resp. $\langle p, q\rangle = \int pq$) when the integral is well defined. We shall consider the $L^2(h)$ norm given by $\|\, p \,\|_{L^2(h)} = \left(\int p^2 h\right)^{1/2}$.

5.3 Convex aggregation for the Kullback-Leibler divergence

In this section, we propose two convex aggregation methods, suited for models submitted to different type of constraints. First, we state non-asymptotic oracle inequalities for the Kullback-Leibler divergence in general form. Then, we derive more explicit non-asymptotic bounds for two applications: the probability density model and the spectral density of stationary Gaussian processes, respectively.

5.3.1 Aggregation procedures

In this section, we describe the two aggregation methods of f using the estimators $(f_k, 1 \leq k \leq N)$. The first one is the convex aggregation of the centered logarithm $(t_k, 1 \leq k \leq N)$ which provides an aggregate estimator f_λ^D. This is particularly useful when considering density estimation, as the final estimator is also a density function. The second one is the convex aggregation of the logarithm $(g_k, 1 \leq k \leq N)$ which provides an aggregate estimator f_λ^S. This method is suitable for spectral density estimation and it can be used for density estimation as well.

Density functions

In this Section, we shall consider probability density function, but what follows can readily be adapted to functions with any given total mass. Notice that if $f \in \mathcal{G}$ is a density, then we get $D(h\|f) = \psi_f$, which in turn implies that $\psi_f \geq 0$ that is, using also the last inequality of (5.8):

$$0 \leq \psi_f \leq \|t_f\|_\infty . \tag{5.11}$$

We want to estimate a density function $f \in \mathcal{G}$ based on the estimators $f_k \in \mathcal{G}$ for $1 \leq k \leq N$ which we assume to be probability density functions. Recall the representation (5.10) of f and f_k with $t = t_f$ and $t_k = t_{f_k}$. For $\lambda \in \Lambda^+$ defined by (5.3), we consider the aggregate estimator f_λ^D given by the convex combination of $(t_k, 1 \leq k \leq N)$:

$$f_\lambda^D = \exp(t_\lambda - \psi_\lambda)\ h \quad \text{with} \quad t_\lambda = \sum_{k=1}^N \lambda_k t_k \quad \text{and} \quad \psi_\lambda = \log\left(\int e^{t_\lambda}\ h\right).$$

Notice that f_λ^D is a density function with $t_{f_\lambda^D} = t_\lambda$, $\psi_{f_\lambda^D} = \psi_\lambda$, and that $\|t_\lambda\|_\infty \leq \max_{1\leq k\leq N} \|t_k\|_\infty < +\infty$, that is $f_\lambda^D \in \mathcal{G}$. The Kullback-Leibler divergence for the estimator f_λ^D of f is given by:

$$D\left(f\|f_\lambda^D\right) = \int f \log\left(f/f_\lambda^D\right) = \langle t - t_\lambda, f\rangle + (\psi_\lambda - \psi). \tag{5.12}$$

Minimizing the Kullback-Leibler distance is thus equivalent to maximizing $\lambda \mapsto \langle t_\lambda, f\rangle - \psi_\lambda$. Notice that $\langle t_\lambda, f\rangle$ is linear in λ and the function $\lambda \mapsto \psi_\lambda$ is convex since $\nabla^2\psi_\lambda$ is the covariance matrix of the random vector $(t_k(Y_\lambda), 1 \leq k \leq N)$ with Y_λ having probability density function f_λ^D. As I_n is a non-negative estimator of f based on the sample $X = (X_1, \ldots, X_n)$, we estimate the scalar product $\langle t_\lambda, f\rangle$ by $\langle t_\lambda, I_n\rangle$. To select the aggregation weights λ, we consider on Λ^+ the penalized empirical criterion $H_n^D(\lambda)$ given by:

$$H_n^D(\lambda) = \langle t_\lambda, I_n\rangle - \psi_\lambda - \frac{1}{2}\,\mathrm{pen}^D(\lambda), \tag{5.13}$$

with penalty term:

$$\mathrm{pen}^D(\lambda) = \sum_{k=1}^N \lambda_k\, D\left(f_\lambda^D\|f_k\right) = \sum_{k=1}^N \lambda_k\psi_k - \psi_\lambda.$$

Remark 5.1. The penalty term in (5.13) can be multiplied by any constant $\theta \in (0,1)$ instead of $1/2$. The choice of $1/2$ is optimal in the sense that it ensures that the constant $\exp(-6K)/4$ in (5.22) of Proposition 5.3 is maximal, giving the sharpest result.

The penalty term is always non-negative and finite. Let $L_n^D(\lambda) = \langle t_\lambda, I_n\rangle - \frac{1}{2}\sum_{k=1}^N \lambda_k\psi_k$. Notice that $L_n^D(\lambda)$ is linear in λ, and that H_n^D simplifies to:

$$H_n^D(\lambda) = L_n^D(\lambda) - \frac{1}{2}\,\psi_\lambda. \tag{5.14}$$

Lemma 5.2 below asserts that the function H_n^D, defined by (5.13), admits a unique maximizer on Λ^+ and that it is strictly concave around this maximizer.

Lemma 5.2. *Let f and $(f_k, 1 \leq k \leq N)$ be density functions, elements of $\mathcal{G}$ such that $(t_k, 1 \leq k \leq N)$ are linearly independent. Then there exists a unique $\hat\lambda_*^D \in \Lambda^+$ such that:*

$$\hat\lambda_*^D = \operatorname*{argmax}_{\lambda\in\Lambda^+}\ H_n^D(\lambda). \tag{5.15}$$

Furthermore, for all $\lambda \in \Lambda^+$, we have:

$$H_n^D(\hat\lambda_*^D) - H_n^D(\lambda) \geq \frac{1}{2}\,D\left(f_{\hat\lambda_*^D}^D\|f_\lambda^D\right). \tag{5.16}$$

Proof. Consider the form (5.14) of $H_n^D(\lambda)$. Recall that the function $\lambda \mapsto L_n^D(\lambda)$ is linear in λ and that $\lambda \mapsto \psi_\lambda$ is convex. Notice that $\nabla\psi_\lambda = (\langle t_k, f_\lambda^D \rangle, 1 \le k \le N)$. This implies that for all $\lambda, \lambda' \in \Lambda^+$:

$$\begin{aligned}(\lambda - \lambda') \cdot \nabla\psi_{\lambda'} + D\left(f_{\lambda'}^D \| f_\lambda^D\right) &= \sum_{k=1}^N (\lambda_k - \lambda'_k) \left\langle t_k, f_{\lambda'}^D \right\rangle + \left\langle t_{\lambda'} - t_\lambda, f_{\lambda'}^D \right\rangle + \psi_\lambda - \psi_{\lambda'} \\ &= \psi_\lambda - \psi_{\lambda'}. \end{aligned} \tag{5.17}$$

Since ψ_λ is convex and differentiable, we deduce from (5.14) that H_n^D is concave and differentiable. We also have by the linearity of L_n^D and (5.17) that for all $\lambda, \lambda' \in \Lambda^+$:

$$H_n^D(\lambda) - H_n^D(\lambda') = (\lambda - \lambda') \cdot \nabla H_n^D(\lambda') - \frac{1}{2} D\left(f_{\lambda'}^D \| f_\lambda^D\right). \tag{5.18}$$

The concave function H_n^D on a compact set attains its maximum at some points $\Lambda^* \subset \Lambda^+$. For $\hat{\lambda}_* \in \Lambda^*$, we have for all $\lambda \in \Lambda^+$:

$$(\lambda - \hat{\lambda}_*) \cdot \nabla H_n^D(\hat{\lambda}_*) \le 0, \tag{5.19}$$

see for example Equation 4.21 of [26]. Using (5.18) with $\lambda' = \hat{\lambda}_*$ and (5.19), we get (5.16) . Let $\hat{\lambda}_*^1$ and $\hat{\lambda}_*^2$ be elements of Λ^*. Then by (5.16), we have:

$$0 = H_n^D(\hat{\lambda}_*^1) - H_n^D(\hat{\lambda}_*^2) \ge \frac{1}{2} D\left(f_{\hat{\lambda}_*^1}^D \| f_{\hat{\lambda}_*^2}^D\right),$$

which implies that a.e. $f_{\hat{\lambda}_*^1}^D = f_{\hat{\lambda}_*^2}^D$. By the linear independence of $(t_k, 1 \le k \le N)$, this gives $\hat{\lambda}_*^1 = \hat{\lambda}_*^2$, giving the uniqueness of the maximizer. □

Using $\hat{\lambda}_*^D$ defined in (5.15), we set:

$$\hat{f}_*^D = f_{\hat{\lambda}_*^D}^D, \quad \hat{t}_*^D = t_{\hat{\lambda}_*^D} \quad \text{and} \quad \hat{\psi}_*^D = \psi_{\hat{\lambda}_*^D}. \tag{5.20}$$

We show that the convex aggregate estimator $\hat{f}_*^D$ verifies almost surely the following non-asymptotic inequality with a bias and a variance term.

Proposition 5.3. *Let $K > 0$. Let f and $(f_k, 1 \le k \le N)$ be probability density functions, elements of $\mathcal{G}$ such that $(t_k, 1 \le k \le N)$ are linearly independent and $\max_{1 \le k \le N} \|t_k\|_\infty \le K$. Let $X = (X_1, \ldots, X_n)$ be a sample from the model P_f. Then the following inequality holds:*

$$D\left(f \| \hat{f}_*^D\right) - D\left(f \| f_{k^*}\right) \le B_n\left(\hat{t}_*^D - t_{k^*}\right) + \max_{1 \le k \le N} V_n^D(e_k),$$

with the functional B_n given by, for $\ell \in L^\infty(\mathbb{R})$:

$$B_n(\ell) = \left\langle \ell, \bar{f}_n - f \right\rangle. \tag{5.21}$$

and the function $V_n^D : \Lambda^+ \to \mathbb{R}$ given by:

$$V_n^D(\lambda) = \left\langle I_n - \bar{f}_n, t_\lambda - t_{k^*} \right\rangle - \frac{\mathrm{e}^{-6K}}{4} \sum_{k=1}^N \lambda_k \|t_k - t_{k^*}\|_{L^2(h)}^2. \tag{5.22}$$

Proof. Using (5.12), we get:

$$D\left(f \| \hat{f}_*^D\right) - D\left(f \| f_{k^*}\right) = \left\langle t_{k^*} - \hat{t}_*^D, f \right\rangle + \hat{\psi}_*^D - \psi_{k^*}.$$

By the definition of k^*, together with $\mathrm{pen}^D(e_k) = 0$ for all $1 \leq k \leq N$ and the strict concavity (5.16) of H_n^D at $\hat{\lambda}_*^D$ with $\lambda = e_{k^*}$, we get:

$$\begin{aligned} D\left(f\|\hat{f}_*^D\right) - D\left(f\|f_{k^*}\right) &\leq \left\langle t_{k^*} - \hat{t}_*^D, f\right\rangle + \hat{\psi}_*^D - \psi_{k^*} + H_n^D(\hat{\lambda}_*^D) - H_n^D(e_{k^*}) - \frac{1}{2} D\left(\hat{f}_*^D\|f_{k^*}\right) \\ &= \left\langle \hat{t}_*^D - t_{k^*}, I_n - f\right\rangle - \frac{1}{2} D\left(\hat{f}_*^D\|f_{k^*}\right) - \frac{1}{2}\mathrm{pen}^D(\hat{\lambda}_*^D) \\ &= B_n\left(\hat{t}_*^D - t_{k^*}\right) + A_n^D, \end{aligned}$$

with:

$$A_n^D = \left\langle \hat{t}_*^D - t_{k^*}, I_n - \bar{f}_n\right\rangle - \frac{1}{2} D\left(\hat{f}_*^D\|f_{k^*}\right) - \frac{1}{2}\sum_{k=1}^N \hat{\lambda}_{*,k}^D D\left(\hat{f}_*^D\|f_k\right). \tag{5.23}$$

We recall, see Lemma 1 of [12], that for any non-negative integrable functions p and q on $\mathbb{R}^d$ satisfying $\|\log(p/q)\|_\infty < +\infty$, we have:

$$D(p\|q) \geq \frac{1}{2}\mathrm{e}^{-\|\log(p/q)\|_\infty}\int p\left(\log(p/q)\right)^2. \tag{5.24}$$

We have:

$$\begin{aligned} D\left(\hat{f}_*^D\|f_k\right) &\geq \frac{1}{2}\mathrm{e}^{-\|\log(\hat{f}_*^D/f_k)\|_\infty}\int \hat{f}_*^D\left(\log(\hat{f}_*^D/f_k)\right)^2 \\ &\geq \frac{1}{2}\mathrm{e}^{-4K-\|\hat{t}_*^D-\hat{\psi}_*^D\|_\infty}\int h\left(\log(\hat{f}_*^D/f_k)\right)^2 \\ &\geq \frac{1}{2}\mathrm{e}^{-6K}\left(\|\hat{t}_*^D - t_k\|_{L^2(h)}^2 + (\hat{\psi}_*^D - \psi_k)^2\right) \\ &\geq \frac{1}{2}\mathrm{e}^{-6K}\|\hat{t}_*^D - t_k\|_{L^2(h)}^2, \end{aligned}$$

where we used (5.24) for the first inequality, (5.11) for the second, and (5.11) as well as $\int t_f h = 0$ for third. By using this lower bound on $D\left(\hat{f}_*^D\|f_k\right)$ to both terms on the right hand side of (5.23), we get:

$$\begin{aligned} A_n^D &\leq \left\langle \hat{t}_*^D - t_{k^*}, I_n - \bar{f}_n\right\rangle - \frac{\mathrm{e}^{-6K}}{4}\|\hat{t}_*^D - t_{k^*}\|_{L^2(h)}^2 - \frac{\mathrm{e}^{-6K}}{4}\sum_{k=1}^N \hat{\lambda}_{*,k}^D\|\hat{t}_*^D - t_k\|_{L^2(h)}^2 \\ &= \left\langle \hat{t}_*^D - t_{k^*}, I_n - \bar{f}_n\right\rangle - \frac{\mathrm{e}^{-6K}}{4}\sum_{k=1}^N \hat{\lambda}_{*,k}^D\|t_k - t_{k^*}\|_{L^2(h)}^2 \\ &= V_n^D(\hat{\lambda}_*^D), \end{aligned}$$

where the first equality is due to the following bias-variance decomposition equality which holds for all $\ell \in L^2(h)$ and $\lambda \in \Lambda^+$:

$$\sum_{k=1}^N \lambda_k\|t_k - \ell\|_{L^2(h)}^2 = \|t_\lambda - \ell\|_{L^2(h)}^2 + \sum_{k=1}^N \lambda_k\|t_\lambda - t_k\|_{L^2(h)}^2. \tag{5.25}$$

The function V_n^D is affine in λ, therefore it takes its maximum on Λ^+ at some e_k, $1 \leq k \leq N$, giving:

$$D\left(f\|\hat{f}_*^D\right) - D\left(f\|f_{k^*}\right) \leq B_n\left(\hat{t}_*^D - t_{k^*}\right) + \max_{1\leq k\leq N} V_n^D(e_k).$$

This concludes the proof. □

Non-negative functions

In this Section, we shall consider non-negative functions. We want to estimate a function $f \in \mathcal{G}$ based on the estimators $f_k \in \mathcal{G}$ for $1 \leq k \leq N$. Since most of the proofs in this Section are similar to those in Section 5.3.1, we only give them when there is a substantial new element. Recall the representation (5.10) of f and f_k. For $\lambda \in \Lambda^+$ defined by (5.3), we consider the aggregate estimator f_λ^D given by the convex aggregation of $(g_k, 1 \leq k \leq N)$:

$$f_\lambda^S = \exp(g_\lambda)\ h \quad \text{with} \quad g_\lambda = \sum_{k=1}^N \lambda_k g_k. \tag{5.26}$$

Notice that $\|g_\lambda\|_\infty \leq \max_{1\leq k\leq N} \|g_k\|_\infty < +\infty$, that is $f_\lambda^D \in \mathcal{G}$. We set $m_\lambda = m_{f_\lambda^S}$ the integral of f_λ^S, see (5.7). The Kullback-Leibler distance for the estimator f_λ^S of f is given by:

$$D\left(f\|f_\lambda^S\right) = \int f \log\left(f/f_\lambda^S\right) - m + m_\lambda = \langle g - g_\lambda, f\rangle - m + m_\lambda. \tag{5.27}$$

Since both g and g_λ are bounded, we deduce that $D\left(f\|f_\lambda^S\right) < \infty$ for all $\lambda \in \Lambda^+$. Minimization of the Kullback-Leibler distance given in (5.27) is therefore equivalent to maximizing $\lambda \mapsto \langle g_\lambda, f\rangle - m_\lambda$. Notice that $\langle g_\lambda, f\rangle$ is linear in λ and the function $\lambda \mapsto m_\lambda$ is convex, since the Hessian matrix $\nabla^2 m_\lambda$ is given by: $\left[\nabla^2 m_\lambda\right]_{i,j} = \int g_i g_j f_\lambda^S$, which is positive-semidefinite. As I_n is a non-negative estimator of f based on the sample $X = (X_1, \ldots, X_n)$, we estimate the scalar product $\langle g_\lambda, f\rangle$ by $\langle g_\lambda, I_n\rangle$. Here we select the aggregation weights λ based on the penalized empirical criterion $H_n^S(\lambda)$ given by:

$$H_n^S(\lambda) = \langle g_\lambda, I_n\rangle - m_\lambda - \frac{1}{2}\,\mathrm{pen}^S(\lambda), \tag{5.28}$$

with the penalty term:

$$\mathrm{pen}^S(\lambda) = \sum_{k=1}^N \lambda_k\, D\left(f_\lambda^S\|f_k\right) = \sum_{k=1}^N \lambda_k m_k - m_\lambda.$$

The choice of the factor $1/2$ for the penalty is justified by arguments similar to those given in Remarks 5.1. The penalty term is always non-negative and finite. Let us define $L_n^S(\lambda) = \langle g_\lambda, I_n\rangle - 1/2 \sum_{k=1}^N \lambda_k m_k$. Notice that $L_n^S(\lambda)$ is linear in λ, and that H_n^S simplifies to:

$$H_n^S(\lambda) = L_n^S(\lambda) - \frac{1}{2}\, m_\lambda. \tag{5.29}$$

Lemma 5.4 below asserts that the function H_n^S admits a unique maximizer on Λ^+ and that it is strictly concave around this maximizer.

Lemma 5.4. *Let f and $(f_k, 1 \leq k \leq N)$ be elements of $\mathcal{G}$ such that $(g_k, 1 \leq k \leq N)$ are linearly independent. Let H_n^S be defined by (5.28). Then there exists a unique $\hat{\lambda}_*^S \in \Lambda^+$ such that:*

$$\hat{\lambda}_*^S = \operatorname*{argmax}_{\lambda\in\Lambda^+}\ H_n^S(\lambda). \tag{5.30}$$

Furthermore, for all $\lambda \in \Lambda^+$, we have:

$$H_n^S(\hat{\lambda}_*^S) - H_n^S(\lambda) \geq \frac{1}{2}\, D\left(f_{\hat{\lambda}_*^S}^S\|f_\lambda^S\right). \tag{5.31}$$

Proof. Notice that for all $\lambda, \lambda' \in \Lambda^+$:

$$m_\lambda - m_{\lambda'} = (\lambda - \lambda')\cdot\nabla m_{\lambda'} + D\left(f_{\lambda'}\|f_\lambda\right). \tag{5.32}$$

The proof is then similar to the proof of Lemma 5.2 using (5.32) instead of (5.17). □

Using $\hat{\lambda}_*^S$ defined in (5.30), we set:

$$\hat{f}_*^S = f_{\hat{\lambda}_*^D}^S \quad \text{and} \quad \hat{g}_*^S = g_{\hat{\lambda}_*^S}. \tag{5.33}$$

We show that the convex aggregate estimator $\hat{f}_*^S$ verifies almost surely the following non-asymptotic inequality with a bias and a variance term.

Proposition 5.5. *Let $K > 0$. Let f and $(f_k, 1 \le k \le N)$ be elements of $\mathcal{G}$ such that $(g_k, 1 \le k \le N)$ are linearly independent and $\max_{1\le k\le N} \| g_k \|_\infty \le K$. Let $X = (X_1, \dots, X_n)$ be a sample from the model P_f. Then the following inequality holds:*

$$D\left(f \| \hat{f}_*^S\right) - D\left(f \| f_{k^*}\right) \le B_n\left(\hat{g}_*^S - g_{k^*}\right) + \max_{1\le k\le N} V_n^S(e_k),$$

with the functional B_n given by (5.21), and the function $V_n^S : \Lambda^+ \to \mathbb{R}$ given by:

$$V_n^S(\lambda) = \left\langle g_\lambda - g_{k^*}, I_n - \bar{f}_n \right\rangle - \frac{\mathrm{e}^{-3K}}{4} \sum_{k=1}^N \lambda_k \| g_k - g_{k^*} \|_{L^2(h)}^2 .$$

Proof. Similarly to the proof of Proposition 5.3 we obtain that:

$$D\left(f \| \hat{f}_*^S\right) - D\left(f \| f_{k^*}\right) \le B_n\left(\hat{g}_*^S - g_{k^*}\right) + A_n^S,$$

with:

$$A_n^S = \left\langle \hat{g}_*^S - g_{k^*}, I_n - \bar{f}_n \right\rangle - \frac{1}{2} D\left(\hat{f}_*^S \| f_{k^*}\right) - \frac{1}{2} \sum_{k=1}^N \hat{\lambda}_{*,k}^S D\left(\hat{f}_*^S \| f_k\right). \tag{5.34}$$

Since $\| \log(\hat{f}_*^S / f_k) \|_\infty = \| g_{\hat{\lambda}^*} - g_k \| \le 2K$ for $1 \le k \le N$, we can apply (5.24) with $\hat{f}_*^S$ and f_k:

$$\begin{aligned} D\left(\hat{f}_*^S \| f_k\right) &\ge \frac{1}{2} \mathrm{e}^{-\left\| \log(\hat{f}_*^S / f_k) \right\|_\infty} \int \hat{f}_*^S \left(\log(\hat{f}_*^S / f_k)\right)^2 \\ &\ge \frac{1}{2} \mathrm{e}^{-2K - \|\hat{g}_*^S\|_\infty} \int h \left(\hat{g}_*^S - g_k\right)^2 \\ &\ge \frac{1}{2} \mathrm{e}^{-3K} \| \hat{g}_*^S - g_k \|_{L^2(h)}^2, \end{aligned} \tag{5.35}$$

where in the second and third inequalities we use that $\| \hat{g}_*^S \|_\infty \le \max_{1\le k\le N} \| g_k \|_\infty \le K$. Applying (5.35) to both terms on the right hand side of (5.34) gives:

$$\begin{aligned} A_n(\hat{\lambda}_*^S) &\le \left\langle \hat{g}_*^S - g_{k^*}, I_n - \bar{f}_n \right\rangle - \frac{\mathrm{e}^{-3K}}{4} \| \hat{g}_*^S - g_{k^*} \|_{L^2(h)}^2 - \frac{\mathrm{e}^{-3K}}{4} \sum_{k=1}^N \hat{\lambda}_{*,k}^S \| \hat{g}_*^S - g_k \|_{L^2(h)}^2 \\ &= \left\langle \hat{g}_*^S - g_{k^*}, I_n - \bar{f}_n \right\rangle - \frac{\mathrm{e}^{-3K}}{4} \sum_{k=1}^N \hat{\lambda}_{*,k}^S \| g_k - g_{k^*} \|_{L^2(h)}^2 \\ &= V_n^S(\hat{\lambda}_*^S), \end{aligned}$$

where we used (5.25) for the second equality. The function V_n^S is affine in λ, therefore it takes its maximum on Λ^+ at some e_k, $1 \le k \le N$, giving:

$$D\left(f \| \hat{f}_*^S\right) - D\left(f \| f_{k^*}\right) \le B_n\left(\hat{g}_*^S - g_{k^*}\right) + \max_{1\le k\le N} V_n^S(e_k).$$

This concludes the proof. □

5.3.2 Applications

In this section we apply the methods established in Section 5.3.1 and 5.3.1 to the problem of density estimation and spectral density estimation, respectively. By construction, the aggregate f_λ^D of Section 5.3.1 is more adapted for the density estimation problem as it produces a proper density function. For the spectral density estimation problem, the aggregate f_λ^S will provide the correct results.

Probability density estimation

We consider the following subset of probability density functions, for $L > 0$:

$$\mathcal{F}^D(L) = \{f \in \mathcal{G};\ \|t_f\|_\infty \leq L \text{ and } m_f = 1\}.$$

The model $\{\mathrm{P}_f, f \in \mathcal{F}^D(L)\}$ corresponds to i.i.d. random sampling from a probability density $f \in \mathcal{F}^D(L)$, that is the random variable $X = (X_1, \ldots . X_n)$ has density $f^{\otimes n}(x) = \prod_{i=1}^n f(x_i)$, with $x = (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$. We estimate the probability measure $f(y)dy$ by the empirical probability measure $I_n(dy)$ given by:

$$I_n(dy) = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}(dy),$$

where δ_y is the Dirac measure at $y \in \mathbb{R}^d$. Notice that I_n is an unbiased estimator of f:

$$f(y)dy = \mathbb{E}[I_n(dy)] \qquad \text{for } y = \mathbb{R}^d.$$

In the following Theorem, we give a sharp non-asymptotic oracle inequality in probability for the aggregation procedure $\hat{f}_*^D$ with a remainder term of order $\log(N)/n$. We prove in Section 5.4.1 the lower bound giving that this remainder term is optimal.

Theorem 5.6. *Let $L, K > 0$. Let $f \in \mathcal{F}^D(L)$ and $(f_k, 1 \leq k \leq N)$ be elements of $\mathcal{F}^D(K)$ such that $(t_k, 1 \leq k \leq N)$ are linearly independent. Let $X = (X_1, \ldots, X_n)$ be an i.i.d. sample from f. Let $\hat{f}_*^D$ be given by (5.20). Then for any $x > 0$ we have with probability greater than $1 - \exp(-x)$:*

$$D\left(f \| \hat{f}_*^D\right) - D\left(f \| f_{k^*}\right) \leq \frac{\beta(\log(N) + x)}{n},$$

with $\beta = 2\exp(6K + 2L) + 4K/3$.

Proof. By Proposition 5.3, we have that:

$$D\left(f \| \hat{f}_*^D\right) - D\left(f \| f_{k^*}\right) \leq B_n\left(\hat{t}_*^D - t_{k^*}\right) + \max_{1 \leq k \leq N} V_n^D(e_k). \tag{5.36}$$

Since $I_n(dy)$ is an unbiased estimator of $f(y)dy$, we get $B_n\left(\hat{t}_*^D - t_{k^*}\right) = 0$. Notice that

$$\mathbb{P}\left(V_n^D(e_k) \geq \frac{\beta(\log(N) + x)}{n}\right) \leq \frac{\mathrm{e}^{-x}}{N} \quad \text{for all } 1 \leq k \leq N, \tag{5.37}$$

implies

$$\mathbb{P}\left(\max_{1 \leq k \leq N} V_n^D(e_k) \geq \frac{\beta(\log(N) + x)}{n}\right) \leq \mathrm{e}^{-x},$$

which will provide a control of the second term on the right hand side of (5.36). Thus, the proof of the theorem will be complete as soon as (5.37) is proved.

To prove (5.37), we use the concentration inequality of Proposition 5.3 in [17] which states that for $Y_1, \ldots, Y_n$ independent random variables with finite variances such that $|Y_i - \mathbb{E}Y_i| \leq b$ for all $1 \leq i \leq n$, we have for all $u > 0$ and $a > 0$:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (Y_i - \mathbb{E}Y_i - a\mathrm{Var}\ Y_i) > \left(\frac{1}{2a} + \frac{b}{3}\right)\frac{u}{n}\right) \leq \mathrm{e}^{-u}. \tag{5.38}$$

Let us choose $Y_i = t_k(X_i) - t_{k^*}(X_i)$ for $1 \leq i \leq n$. Then, since f_k and f_{k^*} belong to $\mathcal{F}^D(K)$, we have $|Y_i - \mathbb{E}Y_i| \leq 4K$, and:

$$\mathrm{Var}\ Y_i \leq \int (t_k - t_{k^*})^2 f \leq \mathrm{e}^{2L}\|t_k - t_{k^*}\|_{L^2(h)}^2. \tag{5.39}$$

Applying (5.38) with $a = \exp(-6K - 2L)/4$, $b = 4K$ and $u = \log(N) + x$, we obtain:

$$\begin{aligned}\frac{\mathrm{e}^{-x}}{N} &\geq \mathbb{P}\left(\left\langle t_k - t_{k^*}, I_n - \bar{f}_n\right\rangle - \frac{\mathrm{e}^{-6K-2L}}{4}\mathrm{Var}\, Y_1 > \frac{\beta(\log(N)+x)}{n}\right)\\ &\geq \mathbb{P}\left(\left\langle t_k - t_{k^*}, I_n - \bar{f}_n\right\rangle - \frac{\mathrm{e}^{-6K}}{4}\,\|t_k - t_{k^*}\|^2_{L^2(h)} > \frac{\beta(\log(N)+x)}{n}\right)\\ &= \mathbb{P}\left(V_n^D(e_k) > \frac{\beta(\log(N)+x)}{n}\right),\end{aligned}$$

where the second inequality is due to (5.39). This proves (5.37) and completes the proof. □

Remark 5.7. We can also use the aggregation method of Section 5.3.1 and consider the normalized estimator $\tilde{f}_*^S = \hat{f}_*^S / m_{\hat{\lambda}_*^S} = f^D_{\hat{\lambda}_*^S}$, which is a proper density function. Notice that the optimal weights $\hat{\lambda}_*^D$ (which defines $\hat{f}_*^D$) and $\hat{\lambda}_*^S$ (which defines $\tilde{f}_*^S$) maximize different criteria. Indeed, according to (5.30) the vector $\hat{\lambda}_*^S$ maximizes:

$$H_n^S(\lambda) = \langle g_\lambda, I_n\rangle - \frac{1}{2}\,m_\lambda - \frac{1}{2}\sum_{k=1}^N \lambda_k m_k = \langle g_\lambda, I_n\rangle - \frac{1}{2}\,m_\lambda - \frac{1}{2},$$

and according to (5.15) the vector $\hat{\lambda}_*^D$ maximizes:

$$H_n^D(\lambda) = \langle t_\lambda, I_n\rangle - \frac{1}{2}\,\psi_\lambda - \frac{1}{2}\sum_{k=1}^N \lambda_k\psi_k = \langle g_\lambda, I_n\rangle - \frac{1}{2}\,\psi_\lambda + \frac{1}{2}\sum_{k=1}^N \lambda_k\psi_k = \langle g_\lambda, I_n\rangle - \frac{1}{2}\log(m_\lambda),$$

where we used the identity $g_\lambda = t_\lambda - \sum_{k=1}^N \lambda_k\psi_k$ for the second equality and the equality $\log(m_\lambda) = \log\left(\int \mathrm{e}^{t_\lambda - \sum_{k=1}^N \lambda_k\psi_k}\, h\right) = \psi_\lambda - \sum_{k=1}^N \lambda_k\psi_k$ for the third.

Spectral density estimation

In this section we apply the convex aggregation scheme of Section 5.3.1 to spectral density estimation of stationary centered Gaussian sequences. Let $h = 1/(2\pi)\mathbf{1}_{[-\pi,\pi]}$ be the reference density and $(X_k)_{k\in\mathbb{Z}}$ be a stationary, centered Gaussian sequence with covariance γ function defined as, for $j \in \mathbb{Z}$:

$$\gamma_j = \mathrm{Cov}\,(X_k, X_{k+j}).$$

Notice that $\gamma_{-j} = \gamma_j$. Then the joint distribution of $X = (X_1, \ldots, X_n)$ is a multivariate, centered Gaussian distribution with covariance matrix $\Sigma_n \in \mathbb{R}^{n\times n}$ given by $[\Sigma_n]_{i,j} = \gamma_{i-j}$ for $1 \leq i, j \leq n$. Notice the sequence $(\gamma_j)_{j\in\mathbb{Z}}$ is semi-definite positive.

We make the following standard assumption on the covariance function γ:

$$\sum_{j=0}^{\infty} |\gamma_j| = C_1 < +\infty. \tag{5.40}$$

The spectral density f associated to the process is the even function defined on $[-\pi, \pi]$ whose Fourier coefficients are γ_j:

$$f(x) = \sum_{j\in\mathbb{Z}} \frac{\gamma_j}{2\pi}\,\mathrm{e}^{ijx} = \frac{\gamma_0}{2\pi} + \frac{1}{\pi}\sum_{j=1}^{\infty} \gamma_j \cos(jx).$$

The first condition in (5.40) ensures that the spectral density is well-defined, continuous and bounded by C_1/π. It is also even and non-negative as $(\gamma_j)_{j\in\mathbb{Z}}$ is semi-definite positive. The function f completely characterizes the model as:

$$\gamma_j = \int_{-\pi}^{\pi} f(x)\,\mathrm{e}^{ijx}\,dx = \int_{-\pi}^{\pi} f(x)\cos(jx)\,dx \quad \text{for } j \in \mathbb{Z}. \tag{5.41}$$

For $\ell \in L^1(h)$, we define the corresponding Toeplitz $T_n(\ell)$ of size $n \times n$ by:

$$[T_n(\ell)]_{j,k} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ell(x)\, \mathrm{e}^{i(j-k)x}\, dx.$$

Notice that $T_n(2\pi f) = \Sigma_n$. Some properties of the Toeplitz matrix $T_n(\ell)$ are collected in Section 5.5.1.

We choose the following estimator of f, for $x \in [-\pi, \pi]$:

$$I_n(x) = \frac{\hat{\gamma}_0}{2\pi} + \frac{1}{\pi} \sum_{j=1}^{n-1} \hat{\gamma}_j \cos(jx),$$

with $(\hat{\gamma}_j, 0 \le j \le n-1)$ the empirical estimates of the correlations $(\gamma_j, 1 \le j \le n-1)$:

$$\hat{\gamma}_j = \frac{1}{n} \sum_{i=1}^{n-j} X_i X_{i+j}. \tag{5.42}$$

The function I_n is a biased estimator, where the bias is due to two different sources: truncation of the infinite sum up to n, and renormalization in (5.42) by n instead of $n-j$ (but it is asymptotically unbiased as n goes to infinity if condition (5.40) is satisfied). The expected value $\bar{f}_n$ of I_n is given by:

$$\bar{f}_n(x) = \sum_{|j|<n} \left(1 - \frac{|j|}{n}\right) \frac{\gamma_j}{2\pi}\, \mathrm{e}^{jx} = \frac{\gamma_0}{2\pi} + \frac{1}{\pi} \sum_{j=1}^{n-1} \frac{(n-j)}{n} \gamma_j \cos(jx).$$

In order to be able to apply Proposition 5.5, we assume that f and the estimators $f_1, \dots, f_N$ of f belongs to $\mathcal{G}$ (they are in particular positive and bounded) and are even functions. In particular the estimators $f_1, \dots, f_N$ and the convex aggregate estimator $\hat{f}_*^S$ defined in (5.33) are proper spectral densities of stationary Gaussian sequences.

Remark 5.8. By choosing $h = 1/(2\pi)\mathbf{1}_{[-\pi,\pi]}$, we restrict our attention to spectral densities that are bounded away from $+\infty$ and 0, see [129] and [27] for the characterization of such spectral densities. Note that we can apply the aggregation procedure to non even functions f_k, $1 \le k \le N$, but the resulting estimator would not be a proper spectral density in that case.

To prove a sharp oracle inequality for the spectral density estimation, since I_n is a biased estimator of f, we shall assume some regularity on the functions f and $f_1, \dots, f_N$ in order to be able to control the bias term. More precisely those conditions will be Sobolev conditions on their logarithm, that is on the functions g and $g_1, \dots, g_N$ defined by (5.6).

For $\ell \in L^2(h)$, the corresponding Fourier coefficients are defined for $k \in \mathbb{Z}$ by $a_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{e}^{-ikx}\, \ell(x)\, dx$. From the Fourier series theory, we deduce that $\sum_{k\in\mathbb{Z}} |a_k|^2 = \|\ell\|^2_{L^2(h)}$ and a.e. $\ell(x) = \sum_{k\in\mathbb{Z}} a_k\, \mathrm{e}^{ikx}$. If furthermore $\sum_{k\in\mathbb{Z}} |a_k|$ is finite, then ℓ is continuous, $\ell(x) = \sum_{k\in\mathbb{Z}} a_k\, \mathrm{e}^{ikx}$ for $x \in [-\pi, \pi]$ and $\|\ell\|_\infty \le \sum_{k\in\mathbb{Z}} |a_k|$.

For $r > 0$, we define the Sobolev norm $\|\ell\|_{2,r}$ of ℓ as:

$$\|\ell\|^2_{2,r} = \|\ell\|^2_{L^2(h)} + \{\ell\}^2_{2,r} \quad \text{with} \quad \{\ell\}^2_{2,r} = \sum_{k\in\mathbb{Z}} |k|^{2r} |a_k|^2.$$

The corresponding Sobolev space is defined by:

$$W_r = \{\ell \in L^2(h);\ \|\ell\|_{2,r} < +\infty\}.$$

For $r > 1/2$, we can bound the supremum norm of ℓ by its Sobolev norm:

$$\|\ell\|_\infty \le \sum_{k\in\mathbb{Z}} |a_k| \le \mathcal{C}_r \{\ell\}_{2,r} \le \mathcal{C}_r \|\ell\|_{2,r}, \tag{5.43}$$

where we used Cauchy-Schwarz inequality for the second inequality with

$$\mathcal{C}_r^2 = \sum_{k\in\mathbb{Z}^*} |k|^{-2r} < +\infty. \tag{5.44}$$

The proof of the following Lemma seems to be part of the folklore, but since we didn't find a proper reference, we give it in Section 5.5.2.

Lemma 5.9. *Let $r > 1/2$, $K > 0$. There exists a finite constant $C(r,K)$ such that for any $g \in W_r$ with $\|g\|_{2,r} \le K$, then we have $\|\exp(g)\|_{2,r} \le C(r,K)$.*

For $r > 1/2$, we consider the following subset of functions:

$$\mathcal{F}_r^S(L) = \{f \in \mathcal{G} : \|g_f\|_{2,r} \le L/\mathcal{C}_r \text{ and } g_f \text{ even}\}. \tag{5.45}$$

For $f \in \mathcal{F}_r^S(L)$, we deduce from (5.43) that g_f is continuous (and bounded by L). This implies that f is a positive, continuous, even function and thus a proper spectral density. Notice that $2\pi\|f\|_\infty \le \exp(L)$. We deduce from (5.41) that $\gamma_k = \int_{-\pi}^{\pi} \mathrm{e}^{-ikx} f(x)\,dx$ and thus:

$$\|f\|_{2,r}^2 = \frac{\gamma_0^2}{4\pi^2} + \frac{1}{2\pi^2}\sum_{k=1}^{\infty}(1+k^{2r})\gamma_k^2.$$

Thus Lemma 5.9 and (5.43) imply also that the covariance function associated to $f \in \mathcal{F}_r^S(L)$ satisfies (5.40). We also get that $\sum_{j=1}^{\infty} j\gamma_j^2 < +\infty$, which is a standard assumption for spectral density estimation.

The following Theorem is the main result of this section.

Theorem 5.10. *Let $r > 1/2$, $K, L > 0$. Let $f \in \mathcal{F}_r^S(L)$ and $(f_k, 1 \le k \le N)$ be elements of $\mathcal{F}_r^S(K)$ such that $(g_k, 1 \le k \le N)$ are linearly independent. Let $X = (X_1, \dots, X_n)$ be a sample of a stationary centered Gaussian sequence with spectral density f. Let $\hat{f}_*^S$ be given by (5.26). Then for any $x > 0$, we have with probability higher than $1 - \exp(-x)$:*

$$D\left(f\|\hat{f}_*^S\right) - D\left(f\|f_{k^*}\right) \le \frac{\beta(\log(N)+x)}{n} + \frac{\alpha}{n},$$

with $\beta = 4(K\,\mathrm{e}^L + \mathrm{e}^{2L+3K})$ and $\alpha = 4KC(r,L)/\mathcal{C}_r$.

Remark 5.11. When the value of γ_0 is given, we shall use the aggregation method of Section 5.3.1 after normalizing the estimators f_k, $1 \le k \le N$ by dividing f_k with $m_k = \int f_k$. The final estimator of f would take the form $\tilde{f}^D_{\hat{\lambda}_*^D} = \gamma_0 f^D_{\hat{\lambda}_*^D}$ and verifies a similar sharp oracle inequality as $\hat{f}_*^S$ (that is without the term α/n of Theorem 5.10). When the value of γ_0 is unknown, it could be estimated empirically by $\hat{\gamma}_0 = \frac{1}{n}\sum_{i=1}^n X_i^2$. Then we could use $\hat{\gamma}_0 f^D_{\hat{\lambda}_*^D}$ to estimate f. However the empirical estimation of γ_0 introduces an error term of order $1/\sqrt{n}$, which leads to a suboptimal remainder term for this aggregation method.

Proof. Using Proposition 5.5 and the notations defined there, we have that:

$$D\left(f\|\hat{f}_*^S\right) - D\left(f\|f_{k^*}\right) \le B_n\left(\hat{g}_*^S - g_{k^*}\right) + \max_{1\le k\le N} V_n^S(e_k). \tag{5.46}$$

First step: Concentration inequality for $\max_{1\le k\le N} V_n^S(e_k)$.

We shall prove that

$$\mathbb{P}\left(\max_{1\le k\le N} V_n^S(e_k) \ge \frac{\beta(\log(N)+x)}{n}\right) \le \mathrm{e}^{-x}. \tag{5.47}$$

It is enough to prove that for each $1 \le k \le N$:

$$\mathbb{P}\left(V_n^S(e_k) \ge \frac{\beta u}{n}\right) \le \mathrm{e}^{-u}. \tag{5.48}$$

Indeed take $u = \log(N) + x$ and the union bound over $1 \leq k \leq N$ to deduce (5.47) from (5.48).

The end of this first step is devoted to the proof of (5.48). Recall definition (5.67) of Toeplitz matrices associated to Fourier coefficients. We express the scalar product $\langle \ell, I_n \rangle$ for $\ell \in \mathbb{L}^\infty([-\pi, \pi])$ in a matrix form:

$$\langle \ell, I_n \rangle = \frac{1}{2\pi n} \sum_{i=1}^{n} \sum_{j=1}^{n} X_i X_j \int_{-\pi}^{\pi} \ell(x) \cos((i-j)x)\, dx = \frac{1}{n} X^T T_n(\ell) X. \tag{5.49}$$

We have the following expression of the covariance matrix of X: $\Sigma_n = 2\pi T_n(f)$. Since f is positive, we get that Σ_n is positive-definite. Set $\xi = \Sigma_n^{-1/2} X$ so that ξ is a centered n-dimensional Gaussian vector whose covariance matrix is the n-dimensional identity matrix. By taking the expected value in (5.49), we obtain:

$$\mathbb{E}\,\langle \ell, I_n \rangle = \left\langle \ell, \bar{f}_n \right\rangle = \frac{1}{n} \operatorname{tr}\left(R_n(\ell)\right),$$

where $\operatorname{tr}(A)$ denotes the trace of the matrix A, and $R_n(\ell) = \Sigma_n^{\frac{1}{2}} T_n(\ell) \Sigma_n^{\frac{1}{2}}$. Therefore the difference $\left\langle \ell, I_n - \bar{f}_n \right\rangle$ takes the form:

$$\left\langle \ell, I_n - \bar{f}_n \right\rangle = \frac{1}{n} \left(\xi^T R_n(\ell) \xi - \operatorname{tr}\left(R_n(\ell)\right) \right).$$

We shall take $\ell = g_k - g_{k^*}$. For this reason, we assume that ℓ is even and $\|\ell\|_\infty \leq 2K$. Let $\eta = (\eta_i, 1 \leq i \leq n)$ denote the eigenvalues of the symmetric matrix $R_n(\ell)$, with η_1 having the largest absolute value. Similarly to Lemma 4.2. of [22], we have that for all $a > 0$:

$$\begin{aligned} \mathrm{e}^{-u} &\geq \mathbb{P}\left(\left\langle \ell, I_n - \bar{f}_n \right\rangle \geq \frac{2|\eta_1| u}{n} + \frac{2\|\eta\| \sqrt{u}}{n} \right) \\ &\geq \mathbb{P}\left(\left\langle \ell, I_n - \bar{f}_n \right\rangle \geq \frac{2|\eta_1| u}{n} + \frac{\|\eta\|^2}{an} + \frac{au}{n} \right), \end{aligned} \tag{5.50}$$

where we used for the second inequality that $2\sqrt{vw} \leq v/a + aw$ for all $v, w, a > 0$. Let us give upper bounds for $|\eta_1|$ and $\|\eta\|^2$. We note $\rho(A)$ for $A \in \mathbb{R}^{n \times n}$ the spectral radius of the matrix A. Then by the well-known properties of the spectral radius, we have that:

$$|\eta_1| = \rho(R_n(\ell)) \leq \rho(\Sigma_n)\rho(T_n(\ell))$$

We deduce from (5.68) that $\rho(\Sigma_n) = \rho(2\pi T_n(f)) \leq 2\pi \|f\|_\infty \leq \exp(L)$ and $\rho(T_n(\ell)) \leq \|\ell\|_\infty \leq 2K$. Therefore we obtain:

$$|\eta_1| \leq 2K\, \mathrm{e}^{L}. \tag{5.51}$$

As for $\|\eta\|^2$, we have:

$$\|\eta\|^2 = \operatorname{tr}\left(R_n^2(\ell)\right) = \operatorname{tr}\left((\Sigma_n T_n(\ell))^2\right) \leq \rho(\Sigma_n)^2 \operatorname{tr}\left(T_n^2(\ell)\right) \leq \mathrm{e}^{2L}\, n \|\ell\|_{L^2(h)}^2, \tag{5.52}$$

where we used (5.69) for the last inequality. Using (5.51) and (5.52) in (5.50) gives:

$$\begin{aligned} \mathrm{e}^{-u} &\geq \mathbb{P}\left(\left\langle \ell, I_n - \bar{f}_n \right\rangle \geq \frac{4K\, \mathrm{e}^{L} u}{n} + \frac{\mathrm{e}^{2L} \|\ell\|_{L^2(h)}^2}{a} + \frac{au}{n} \right) \\ &\geq \mathbb{P}\left(\left\langle \ell, I_n - \bar{f}_n \right\rangle - \frac{\mathrm{e}^{-3K}}{4} \|\ell\|_{L^2(h)}^2 \geq \frac{\beta u}{n} \right), \end{aligned}$$

where for the second inequality we set $a = 4\exp(2L + 3K)$. This proves (5.48), thus (5.47).

Second step: Upper bound for the bias term $B_n(\hat{g}_*^S - g_{k^*})$

We set $\ell_* = \hat{g}_*^S - g_{k^*}$ and we have $\|\ell_*\|_{2,r} \leq 2K/\mathcal{C}_r$. Let $(a_k)_{k\in\mathbb{Z}}$ be the corresponding Fourier coefficients, which are real as ℓ_* is even. We decompose the the bias term as follows:

$$B_n(\ell_*) = \left\langle \bar{f}_n - f, \ell_* \right\rangle = \left\langle \bar{f}_{n,1} - f, \ell_* \right\rangle - \left\langle \bar{f}_{n,2}, \ell_* \right\rangle, \tag{5.53}$$

with $\bar{f}_{n,1}, \bar{f}_{n,2}$ given by, for $x \in [-\pi, \pi]$:

$$\bar{f}_{n,1}(x) = \sum_{|j|<n} \frac{\gamma_j}{2\pi} \mathrm{e}^{ijx} \quad \text{and} \quad \bar{f}_{n,2}(x) = \frac{1}{n} \sum_{|j|<n} \frac{|j|\gamma_j}{2\pi} \mathrm{e}^{ijx}.$$

For the first term of the right hand side of (5.53) notice that:

$$\bar{f}_{n,1}(x) - f(x) = -\sum_{|j|\geq n} \frac{\gamma_j}{2\pi} \mathrm{e}^{ijx}.$$

We deduce that $\left\langle \bar{f}_{n,1} - f, \ell_* \right\rangle = \left\langle \bar{f}_{n,1} - f, \bar{\ell}_* \right\rangle$, with $\bar{\ell}_* = \sum_{|j|\geq n} a_j \mathrm{e}^{ijx}$. Then, by the Cauchy-Schwarz inequality, we get:

$$\left|\left\langle \bar{f}_{n,1} - f, \bar{\ell}_* \right\rangle\right| \leq \|\bar{f}_{n,1} - f\|_{L^2(h)} \|\bar{\ell}_*\|_{L^2(h)}.$$

Thanks to Lemma 5.9, we get:

$$\|\bar{f}_{n,1} - f\|_{L^2(h)}^2 = \sum_{|j|\geq n}^{\infty} \frac{\gamma_j^2}{4\pi^2} \leq \sum_{|j|\geq n}^{\infty} \frac{|j|^{2r}}{n^{2r}} \frac{\gamma_j^2}{4\pi^2} \leq \frac{1}{n^{2r}} \{f\}_{2,r}^2 \leq \frac{1}{n^{2r}} \|f\|_{2,r}^2 \leq \frac{C(r,L)^2}{n^{2r}}.$$

This gives $\|\bar{f}_{n,1} - f\|_{L^2(h)} \leq C(r,L) n^{-r}$. Similarly, we have:

$$\|\bar{\ell}_*\|_{L^2(h)} \leq n^{-r} \{\ell_*\}_{2,r} \leq n^{-r} \|\ell_*\|_{2,r} \leq 2Kn^{-r}/\mathcal{C}_r.$$

We deduce that:

$$\left|\left\langle \bar{f}_{n,1} - f, \bar{\ell}_* \right\rangle\right| \leq \frac{2KC(r,L)}{\mathcal{C}_r} n^{-2r}. \tag{5.54}$$

For the second term on the right hand side of (5.53), we have:

$$\left\langle \bar{f}_{n,2}, \ell_* \right\rangle = \frac{1}{n} \sum_{|j|<n} \frac{|j|\gamma_j}{2\pi} a_j.$$

Using the Cauchy-Schwarz inequality and then Lemma 5.9, we get as $r > 1/2$:

$$\left|\left\langle \bar{f}_{n,2}, \ell_* \right\rangle\right| \leq \frac{1}{n} \{\ell_*\}_{2,1/2} \{f\}_{2,1/2} \leq \frac{1}{n} \|\ell_*\|_{2,r} \|f\|_{2,r} \leq \frac{2KC(r,L)}{\mathcal{C}_r} n^{-1}. \tag{5.55}$$

Therefore combining (5.54) and (5.55), we obtain the following upper bound for the bias:

$$|B_n(\ell_*)| \leq \frac{4KC(r,L)}{\mathcal{C}_r} n^{-1}. \tag{5.56}$$

Third step: Conclusion

Use (5.47) and (5.56) in (5.46) to get the result. □

5.4 Lower bounds

In this section we show that the aggregation procedure given in Section 5.3 is optimal by giving a lower bound corresponding to the upper bound of Theorem 5.6 and 5.10 for the estimation of the probability density function as well as for the spectral density.

5.4.1 Probability density estimation

In this section we suppose that the reference density is the uniform distribution on $[0,1]^d$: $h = \mathbf{1}_{[0,1]^d}$.

Remark 5.12. If the reference density is not the uniform distribution on $[0,1]^d$, then we can apply the Rosenblatt transformation, see [151], to reduce the problem to this latter case. More precisely, according to [151], if the random variable Z has probability density h, then there exists two maps T and T^{-1} such that $U = T(Z)$ is uniform on $[0,1]^d$ and a.s. $Z = T^{-1}(U)$. Then if the random variable X has density $f = \exp(g)\,h$, we deduce that $T(X)$ has density $f^T = \exp(g \circ T^{-1})\mathbf{1}_{[0,1]^d}$. Furthermore, if f_1 and f_2 are two densities (with respect to the reference density h), then we have $D(f_1 \| f_2) = D\left(f_1^T \| f_2^T\right)$.

We give the main result of this Section. Let $\mathbb{P}_f$ denote the probability measure when $X_1, \ldots, X_n$ are i.i.d. random variable with density f.

Proposition 5.13. *Let $N \geq 2$, $L > 0$. Then there exist N probability densities $(f_k, 1 \leq k \leq N)$, with $f_k \in \mathcal{F}^D(L)$ such that for all $n \geq 1$, $x \in \mathbb{R}^+$ satisfying:*

$$\frac{\log(N) + x}{n} < 3\left(1 - \mathrm{e}^{-L}\right)^2, \tag{5.57}$$

we have:

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}^D(L)} \mathbb{P}_f\left(D\left(f \| \hat{f}_n\right) - \min_{1 \leq k \leq N} D(f \| f_k) \geq \frac{\beta'\left(\log(N) + x\right)}{n}\right) \geq \frac{1}{24}\,\mathrm{e}^{-x},$$

with the infimum taken over all estimators $\hat{f}_n$ based on the sample $X_1, \ldots, X_n$, and $\beta' = 2^{-17/2}/3$.

In the following proof, we shall use the Hellinger distance which is defined as follows. For two non-negative integrable functions p and q, the Hellinger distance $H(p,q)$ is defined as:

$$H(p,q) = \sqrt{\int (\sqrt{p} - \sqrt{q})^2}.$$

A well known property of this distance is that its square is smaller then the Kullback-Leibler divergence defined by 5.4, that is for all non-negative integrable functions p and q, we have:

$$H^2(p,q) \leq D(p \| q).$$

Proof. Since the probability densities $(f_k, 1 \leq k \leq N)$ belongs to $\mathcal{F}^D(L)$, we have:

$$\begin{aligned}
\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}^D(L)} \mathbb{P}_f\left(D\left(f \| \hat{f}_n\right) - \min_{1 \leq k \leq N} D(f \| f_k) \geq \frac{\beta'\left(\log(N) + x\right)}{n}\right) & \\
\geq \inf_{\hat{f}_n} \max_{1 \leq k \leq N} \mathbb{P}_{f_k}\left(D\left(f_k \| \hat{f}_n\right) \geq \frac{\beta'\left(\log(N) + x\right)}{n}\right) & \\
\geq \inf_{\hat{f}_n} \max_{1 \leq k \leq N} \mathbb{P}_{f_k}\left(H^2(f_k, \hat{f}_n) \geq \frac{\beta'\left(\log(N) + x\right)}{n}\right). &
\end{aligned}$$

For the choice of $(f_k, 1 \leq k \leq N)$, we follow the choice given in the proof of Theorem 2 of [120]. Let D be the smallest positive integer such that $2^{D/8} \geq N$ and $\Delta = \{0,1\}^D$. For $0 \leq j \leq D-1$, $s \in \mathbb{R}$, we set:

$$\alpha_j(s) = \frac{T}{D}\mathbf{1}_{(0,\frac{1}{2}]}(Ds - j) - \frac{T}{D}\mathbf{1}_{(\frac{1}{2},1]}(Ds - j),$$

where T verifies $0 < T \leq D(1 - \mathrm{e}^{-L})$. Notice the support of the function α_j is $(j/D, (j+1)/D]$. Then for any $\delta = (\delta_1, \ldots, \delta_D) \in \Delta$, the function f^δ defined by:

$$f^\delta(y) = 1 + \sum_{j=0}^{D-1} \delta_j \alpha_j(y_1), \quad y = (y_1, \ldots, y_d) \in [0,1]^d,$$

is a probability density function with $\mathrm{e}^L \geq 1 + T/D \geq f \geq 1 - T/D \geq \mathrm{e}^{-L}$. This implies that $f^\delta \in \mathcal{F}^D(L)$. As shown in the proof of Theorem 2 in [120], there exists N probability densities $(f_k, 1 \leq k \leq N)$ amongst $\{f^\delta, \delta \in \Delta\}$ such that for any $i \neq j$, we have:

$$H^2(f_i, f_j) \geq \frac{8^{-3/2} T^2}{4D^2},$$

and f_1 can be chosen to be the density of the uniform distribution on $[0,1]^d$. Recall the notation $p^{\otimes n}$ of the n-product probability density corresponding to the probability density p. Then we also have (see the proof of Theorem 2 of [120]) for all $1 \leq i \leq N$:

$$D\left(f_i^{\otimes n} \| f_1^{\otimes n}\right) \leq \frac{nT^2}{D^2}.$$

Let us take $T = D\sqrt{(\log(N) + x)/3n}$, so that with condition (5.57) we indeed have $T \leq D(1 - \mathrm{e}^{-L})$. With this choice, and the defintion of β', we have for $1 \leq i \neq j \leq N$

$$H^2(f_i, f_j) \geq 4\frac{\beta'(\log(N) + x)}{n} \quad \text{and} \quad D\left(f_i^{\otimes n} \| f_1^{\otimes n}\right) \leq \frac{\log(N) + x}{3}.$$

Now we apply Corollary 5.1 of [17] with $m = N - 1$ and with the squared Hellinger distance instead of the L^2 distance to get that for any estimator $\hat{f}_n$:

$$\max_{1 \leq k \leq N} \mathbb{P}_{f_k}\left(H^2(f_k, \hat{f}_n) \geq \frac{\beta'(\log(N) + x)}{n}\right) \geq \frac{1}{12} \min\left(1, (N-1)\,\mathrm{e}^{-(\log(N)+x)}\right) \geq \frac{1}{24}\,\mathrm{e}^{-x}.$$

This concludes the proof. □

5.4.2 Spectral density estimation

In this section we give a lower bound for aggregation of spectral density estimators. Let $\mathbb{P}_f$ denote the probability measure when $(X_n)_{n \in \mathbb{Z}}$ is a centered Gaussian sequence with spectral density f. Recall the set of positive even function $\mathcal{F}_r^S(L) \subset \mathcal{G}$ defined by (5.45) for $r \in \mathbb{R}$.

Proposition 5.14. *Let $N \geq 2$, $r > 1/2$, $L > 0$. There exist a constant $C(r, L)$ and N spectral densities $(f_k, 1 \leq k \leq N)$ belonging to $\mathcal{F}_r^S(L)$ such that for all $n \geq 1$, $x \in \mathbb{R}^+$ satisfying:*

$$\frac{\log(N) + x}{n} < \frac{C(r, L)}{\log(N)^{2r}} \tag{5.58}$$

we have:

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}_r^S(L)} \mathbb{P}_f\left(D\left(f \| \hat{f}_n\right) - \min_{1 \leq k \leq N} D\left(f \| f_k\right) \geq \frac{\beta'(\log(N) + x)}{n}\right) \geq \frac{1}{24}\,\mathrm{e}^{-x}, \tag{5.59}$$

with the infimum taken over all estimators $\hat{f}_n$ based on the sample sequence $X = (X_1, \ldots, X_n)$, and $\beta' = 8^{-5/2}/3$.

Proof. Similarly to the proof of Proposition 5.13, the left hand side of (5.59) is greater than:

$$\inf_{\hat{f}_n} \max_{1 \leq k \leq N} \mathbb{P}_{f_k}\left(H^2(f_k, \hat{f}_n) \geq \frac{\beta'(\log(N) + x)}{n}\right).$$

We shall choose a set of spectral densities $(f_k, 1 \le k \le N)$ similarly as in the proof of Proposition 5.13 such that $f_k \in \mathcal{F}_r^S(L)$. Let us define $\varphi : [0,\pi] \to \mathbb{R}$ as, for $x \in [0,\pi]$:

$$\varphi(x) = \zeta(x)\mathbf{1}_{[0,\pi/2]}(x) - \zeta(x)\mathbf{1}_{[\pi/2,\pi]}(x) \quad \text{with} \quad \zeta(x) = \mathrm{e}^{-1/x\left(\frac{\pi}{2}-x\right)}.$$

We have that $\varphi \in C^\infty(\mathbb{R})$ and:

$$\|\varphi\|_\infty = \mathrm{e}^{-16/\pi^2}, \quad \int_0^\pi \varphi = 0. \tag{5.60}$$

Let D be the smallest integer such that $2^{D/8} \ge N$ and $\Delta = \{0,1\}^D$. For $1 \le j \le D$, $x \in [0,\pi]$, let $\bar{\alpha}_j(x)$ be defined as:

$$\bar{\alpha}_j(x) = \varphi(Dx - (j-1)\pi),$$

and for any $\delta = (\delta_1, \dots, \delta_D) \in \Delta$ and $s \ge 0$, let the function f_s^δ be defined by:

$$2\pi\, f_s^\delta(y) = 1 + s\sum_{j=1}^D \delta_j \bar{\alpha}_j(|y|), \quad y \in [-\pi,\pi]. \tag{5.61}$$

Since $\int_0^\pi \varphi = 0$, we get:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} f_s^\delta(x)\,dx = 1 \quad \text{and} \quad 1 - s\,\|\varphi\|_\infty \le 2\pi f_s^\delta \le 1 + s\,\|\varphi\|_\infty. \tag{5.62}$$

We assume that $s \in [0,1/2]$, so that $2\pi f_s^\delta \ge 1/2$. Let us denote $g_s^\delta = g_{f_s^\delta} = \log(2\pi f_s^\delta)$. We first give upper bounds for $\|(g_s^\delta)^{(p)}\|_{L^2(h)}$ with $p \in \mathbb{N}$.

For $p = 0$, we have by (5.62) :

$$\|g_s^\delta\|_{L^2(h)} \le \log\left(\frac{1}{1 - s\,\|\varphi\|_\infty}\right) \le \frac{s\,\|\varphi\|_\infty}{1 - s\,\|\varphi\|_\infty} \le 2s. \tag{5.63}$$

For $p \ge 1$, we get by Faà di Bruno's formula that:

$$\|(g_s^\delta)^{(p)}\|_{L^2(h)} = \left\| \sum_{k \in \mathcal{K}_p} \frac{p!}{k_1!k_2!\dots k_p!} \frac{(-1)^{\bar{k}+1}\bar{k}!}{(2\pi f_s^\delta)^{\bar{k}}} \prod_{\ell=1}^p \left(\frac{(2\pi f_s^\delta)^{(\ell)}}{\ell!}\right)^{k_\ell} \right\|_{L^2(h)}, \tag{5.64}$$

with $\mathcal{K}_p = \{k = (k_1,\dots,k_p) \in \mathbb{N}^p; \sum_{\ell=1}^p \ell k_\ell = p\}$ and $\bar{k} = \sum_{\ell=1}^p k_\ell$. The ℓ-th derivative of $2\pi f_s^\delta$ is given by, for $y \in [0,\pi]$:

$$(2\pi f_s^\delta(y))^{(\ell)} = sD^\ell \sum_{j=1}^D \delta_j \varphi^{(\ell)}(Dy - (j-1)\pi).$$

Therefore we have the following bound for this derivative:

$$\|(2\pi f_s^\delta(y))^{(\ell)}\|_\infty \le sD^\ell\,\|\varphi^{(\ell)}\|_\infty.$$

From $\varphi \in C^\infty(\mathbb{R})$, we deduce that $\|\varphi^{(\ell)}\|_\infty$ is finite for all $\ell \in \mathbb{N}^*$. Since $s \in [0,1/2]$ and $2\pi f_s^\delta \ge 1 - s\,\|\varphi\|_\infty \ge 1/2$, there exists a constant $\bar{C}_p$ depending on p (and not depending on N), such that :

$$\|(g_s^\delta)^{(p)}\|_{L^2(h)} \le s\bar{C}_p D^p \le s\bar{C}_p \frac{16^p}{\log(2)^p}\log(N)^p. \tag{5.65}$$

In order to have $f_s^\delta \in \mathcal{F}_r^S(L)$, we need to ensure that $\|g_s^\delta\|_{2,r} \le L/\mathcal{C}_r$. For $r \in \mathbb{N}^*$, we have:

$$\|g_s^\delta\|_{2,r} = \sqrt{\|g_s^\delta\|_{L^2(h)}^2 + \|(g_s^\delta)^{(r)}\|_{L^2(h)}^2}.$$

Therefore if $s \in [0, s_{r,L}]$ with $s_{r,L} \in [0, 1/2]$ given by:

$$s_{r,L} = \log(N)^{-r} \bar{C}_{r,L}, \quad \text{with} \quad \bar{C}_{r,L} = \min\left(\frac{\log(2)^r}{2}, \frac{\log(2)^r L}{\sqrt{8}\mathcal{C}_r}, \frac{\log(2)^r L}{\sqrt{2}\mathcal{C}_r 16^r \bar{C}_r}\right),$$

then by (5.63) and (5.65) we get:

$$\| g_s^\delta \|_{2,r} \leq \sqrt{\frac{L^2}{2\mathcal{C}_r^2} + \frac{L^2}{2\mathcal{C}_r^2}} = \frac{L}{\mathcal{C}_r}.$$

Let $\lceil r \rceil$ and $\lfloor r \rfloor$ denote the unique integers such that $\lceil r \rceil - 1 < r \leq \lceil r \rceil$ and $\lfloor r \rfloor \leq r < \lfloor r \rfloor + 1$. For $r \notin \mathbb{N}^*$, Hölder's inequality yields:

$$\begin{aligned}
\| g_s^\delta \|_{2,r} &= \sqrt{\| g_s^\delta \|_{L^2(h)}^2 + \{g_s^\delta\}_{2,r}^2} \\
&\leq \sqrt{\| g_s^\delta \|_{L^2(h)}^2 + \{g_s^\delta\}_{2,\lceil r \rceil}^{2(r-\lfloor r \rfloor)} \{g_s^\delta\}_{2,\lfloor r \rfloor}^{2(\lceil r \rceil - r)}} \\
&= \sqrt{\| g_s^\delta \|_{L^2(h)}^2 + \| (g_s^\delta)^{(\lceil r \rceil)} \|_{L^2(h)}^{2(r-\lfloor r \rfloor)} \| (g_s^\delta)^{(\lfloor r \rfloor)} \|_{L^2(h)}^{2(\lceil r \rceil - r)}}.
\end{aligned}$$

Using (5.65) and (5.65) with $p = \lceil r \rceil$ and $p = \lfloor r \rfloor$, we obtain:

$$\| (g_s^\delta)^{(\lceil r \rceil)} \|_{L^2(h)}^{2(r-\lfloor r \rfloor)} \| (g_s^\delta)^{(\lfloor r \rfloor)} \|_{L^2(h)}^{2(\lceil r \rceil - r)} \leq s^2 \bar{C}_{\lceil r \rceil}^{2(r-\lfloor r \rfloor)} \bar{C}_{\lfloor r \rfloor}^{2(\lceil r \rceil - r)} \frac{16^{2r}}{\log(2)^{2r}} \log N^{2r}.$$

Hence if $s \in [0, s_{r,L}]$ with $s_{r,L} \in [0, 1/2]$ given by:

$$s_{r,L} = \log(N)^{-r} \bar{C}_{r,L}, \quad \text{with} \quad \bar{C}_{r,L} = \min\left(\frac{\log(2)^r}{2}, \frac{\log(2)^r L}{\sqrt{8}\mathcal{C}_r}, \frac{\log(2)^r L}{\sqrt{2}\mathcal{C}_r 16^r \bar{C}_{\lceil r \rceil}^{r-\lfloor r \rfloor} \bar{C}_{\lfloor r \rfloor}^{\lceil r \rceil - r}}\right),$$

we also have $\| g_s^\delta \|_{2,r} \leq L/\mathcal{C}_r$, providing $f_s^\delta \in \mathcal{F}_r^S(L)$.

Mimicking the proof of Theorem 2 in [120] and omitting the details, we first obtain (see last inequality of p.975 in [120]) that for $\delta, \delta' \in \Delta$:

$$H^2\left(f_s^\delta, f_s^{\delta'}\right) \geq 8^{-3/2} \frac{\sigma(\delta, \delta')}{D} \frac{2}{\pi} s^2 \int_0^\pi \varphi^2,$$

with $\sigma(\delta, \delta')$ the Hamming distance between δ and δ', and then deduce that there exist $(\delta^k, 1 \leq k \leq N)$ in Δ with $\delta^1 = 0$ such that for any $1 \leq i \neq j \leq N$ and $s \in [0, s_{r,L}]$, we have (see first inequality of p.976 in [120]):

$$H^2(f_s^{\delta^i}, f_s^{\delta^j}) \geq \frac{2 \cdot 8^{-5/2}}{\pi} s^2 \int_0^\pi \varphi^2.$$

Notice $f_s^{\delta^1} = f_s^0 = h$ is the density of the uniform distribution on $[-\pi, \pi]$.

With a slight abuse of notation, let us denote by P_f the joint probability density of the centered Gaussian sequence $X = (X_1, \ldots, X_n)$ corresponding to the spectral density f. Assume X is standardized (that is Var $(X_1) = 1$), which implies $\int f = 1$. Let $\Sigma_{n,f}$ denote the corresponding covariance matrix. Since $h = (1/2\pi)\mathbf{1}_{[-\pi,\pi]}$, we have $\Sigma_{n,h} = \mathcal{I}_n$ the $n \times n$-dimensional identity matrix. We compute:

$$\begin{aligned}
D\left(\mathrm{P}_f \| \mathrm{P}_h\right) &= \int_{\mathbb{R}^n} \mathrm{P}_f(x) \log\left(\frac{\mathrm{P}_f(x)}{\mathrm{P}_h(x)}\right) dx \\
&= \int_{\mathbb{R}^n} \mathrm{P}_f(x) \log\left(\frac{1}{\sqrt{\det(\Sigma_{n,f})}} \exp\left(-\frac{1}{2} x^T \left(\Sigma_{n,f}^{-1} - \mathcal{I}_n\right) x\right)\right) dx \\
&= -\frac{1}{2} \log\left(\det(\Sigma_{n,f})\right) - \frac{1}{2} \mathrm{E}_f\left[X^T \left(\Sigma_\delta^{-1} - \mathcal{I}_n\right) X\right].
\end{aligned}$$

The expected value in the previous equality can be written as:

$$\mathrm{E}_f\left[X^T\left(\Sigma_{n,f}^{-1}-\mathcal{I}_n\right)X\right]=\mathrm{tr}\left(\left(\Sigma_{n,f}^{-1}-\mathcal{I}_n\right)\mathrm{E}_f[X^TX]\right)=\mathrm{tr}\left(\mathcal{I}_n-\Sigma_{n,f}\right)=0,$$

where for the last equality, we used that the Gaussian random variables are standardized. This yields $D\left(\mathrm{P}_f\|\mathrm{P}_h\right)=-\frac{1}{2}\log\left(\det(\Sigma_{n,f})\right)$. We can use this last equality for $f=f_s^\delta$ since $\int f_s^\delta=1$ thanks to (5.60), and obtain:

$$D\left(\mathrm{P}_{f_s^\delta}\|\mathrm{P}_{f_s^0}\right)=-\frac{1}{2}\log\left(\det(\Sigma_{n,f_s^\delta})\right).$$

Notice that for $s\in[0,s_{r,L}]$, we have $3/2\geq 1+s\,\|\varphi\|_\infty\geq 2\pi f_s^\delta\geq 1-s\,\|\varphi\|_\infty\geq 1/2$ thanks to (5.62) and (5.60). Therefore we have:

$$D\left(\mathrm{P}_{f_s^\delta}\|\mathrm{P}_{f_s^0}\right)\leq\frac{n}{2}\,\|2\pi f_s^\delta-1\|_{L^2(h)}^2\leq\frac{n}{2}\frac{s^2}{\pi}\int_0^\pi\varphi^2,\tag{5.66}$$

where we used $\Sigma_{n,f_s^\delta}=T_n(2\pi f_s^\delta)$ and Lemma 5.16 with $\ell=2\pi f_s^\delta$ for the first inequality, and (5.61) for the second inequality. We set:

$$C(r,L)=\frac{3\bar{C}_{r,L}^2\int_0^\pi\varphi^2}{2\pi}\quad\text{and}\quad s=\sqrt{\frac{2\pi}{3\int_0^\pi\varphi^2}}\sqrt{\frac{\log(N)+x}{n}},$$

so that (5.58) holds for $s\in[0,s_{r,L}]$. We obtain for all $\delta^1,\delta^2\in\bar{\Delta}$, $\delta\in\Delta$:

$$H^2\left(f_s^{\delta_1},f_s^{\delta_2}\right)\geq 4\frac{\beta'(\log(N)+x)}{n}\quad\text{and}\quad D\left(\mathrm{P}_{f_s^\delta}\|\mathrm{P}_{f_s^0}\right)\leq\frac{\log(N)+x}{3}.$$

We conclude the proof as in the end of the proof of Proposition 5.13.

□

5.5 Appendix

5.5.1 Results on Toeplitz matrices

Let $\ell\in L^1(h)$ be a real function with $h=1/(2\pi)\mathbf{1}_{[-\pi,\pi]}$. We define the corresponding Toeplitz matrix $T_n(\ell)$ of size $n\times n$ of its Fourier coefficients by:

$$[T_n(\ell)]_{j,k}=\frac{1}{2\pi}\int_{-\pi}^{\pi}\ell(x)\,\mathrm{e}^{i(j-k)x}\,dx\quad\text{for } 1\leq j,k\leq n.\tag{5.67}$$

Notice that $T_n(\ell)$ is Hermitian. It is also real if ℓ is even. Recall that $\rho(A)$ denotes the spectral density of the matrix A.

Lemma 5.15. *Let $\ell\in L^2(h)$ be a real function.*

1. *All the eigenvalues of $T_n(\ell)$ belong to $[\min\ell,\max\ell]$. In particular, we have the following upper bound on the spectral radius $\rho(T_n(\ell))$ of $T_n(\ell)$:*

$$\rho(T_n(\ell))\leq\|\ell\|_\infty.\tag{5.68}$$

2. *For the trace of $T_n(\ell)$ and $T_n^2(\ell)$, we have:*

$$\mathrm{tr}\left(T_n(\ell)\right)=\frac{n}{2\pi}\int_{-\pi}^{\pi}\ell(x)\,dx\quad\text{and}\quad\mathrm{tr}\left(T_n^2(\ell)\right)\leq n\,\|\ell\|_{L^2(h)}^2.\tag{5.69}$$

Proof. For Property (1), see Equation (6) of Section 5.2 in [88]. For Property (2), the first part is clear and for the second part, see Lemma 3.1 of [56].

□

We shall use the following elementary result.

Lemma 5.16. *Let $\ell \in L^2(h)$ such that $\int \ell h = 1$ and $\ell(x) \in [1/2, 3/2]$, then we have:*

$$\log\left(\det(T_n(\ell))\right) \geq -n \,\|\, \ell - 1 \,\|^2_{L^2(h)} \,. \tag{5.70}$$

Proof. Notice that by Property (1), the eigenvalues $(\nu_i, 1 \leq i \leq n)$ of $T_n(\ell)$ verify $\nu_i \in [1/2, 3/2]$. For $t \in [-1/2, 1/2]$, we have $\log(1+t) \geq t - t^2$, giving that:

$$\log\left(\det(T_n(\ell))\right) = \sum_{i=1}^{n} \log(\nu_i) \geq \sum_{i=1}^{n} (\nu_i - 1) - (\nu_i - 1)^2 = -\operatorname{tr}\left(T_n^2(\ell - 1)\right) \geq -n \,\|\, \ell - 1 \,\|^2_{L^2(h)},$$

where we used that $T_n(\ell - 1) = T_n(\ell) - \mathcal{I}_n$ for the second equality and Property (2) for the second inequality. □

5.5.2 Proof of Lemma 5.9

The next Lemma is inspired by the work of [64] on fractional Sobolev spaces. For $r \in (0,1)$ and $\ell \in L^2(h)$, we define:

$$I_r(\ell) = \frac{1}{2\pi} \int_{[-\pi,\pi]^2} \frac{|\ell(x+y) - \ell(x)|^2}{|y|^{1+2r}} \, dxdy,$$

where we set $\ell(z) = \ell(z - 2\pi)$ for $z \in (\pi, 2\pi]$ and $\ell(z) = \ell(z + 2\pi)$ for $z \in [-2\pi, -\pi)$.

Lemma 5.17. *Let $r \in (0,1)$ and $\ell \in L^2(h)$. Then we have:*

$$c_r \{\ell\}^2_{2,r} \leq I_r(\ell) \leq C_r \{\ell\}^2_{2,r}. \tag{5.71}$$

Proof. Using the Fourier representation of ℓ, we get:

$$I_r(\ell) = \sum_{k \in \mathbb{Z}} |a_k|^2 \int_{-\pi}^{\pi} \frac{|1 - \mathrm{e}^{iky}|^2}{|y|^{1+2r}} \, dy = \sum_{k \in \mathbb{Z}} |k|^{2r} |a_k|^2 \int_{-|k|\pi}^{|k|\pi} \frac{|1 - \mathrm{e}^{iz}|^2}{|z|^{1+2r}} \, dz.$$

For $r \in (0,1)$ and $k \in \mathbb{Z}^*$, we have

$$0 < c_r := \int_{-\pi}^{\pi} \frac{|1 - \mathrm{e}^{iz}|^2}{|z|^{1+2r}} \, dz \leq \int_{-|k|\pi}^{|k|\pi} \frac{|1 - \mathrm{e}^{iz}|^2}{|z|^{1+2r}} \, dz \leq \int_{\mathbb{R}} \frac{|1 - \mathrm{e}^{iz}|^2}{|z|^{1+2r}} \, dz =: C_r < +\infty.$$

This yields (5.71).

□

First step : $r \in (1/2, 1)$

Let $r \in (1/2, 1)$ and set $L = \mathcal{C}_r K$. Let $f = \mathrm{e}^g$ with $g \in W_r$ such that $\|\, g \,\|_{2,r} \leq K$. Thanks to (5.43), we have $\|\, g \,\|_\infty \leq \mathcal{C}_r K = L$. Using that $|\, \mathrm{e}^x - \mathrm{e}^y \,| \leq \mathrm{e}^L \,|x - y|$ for $x, y \in [-L, L]$, we deduce that:

$$I_r(f) = I_r(\mathrm{e}^g) \leq \mathrm{e}^{2L} \, I_r(g) \quad \text{and} \quad \|\, f \,\|^2_{L^2(h)} \leq \mathrm{e}^{2L} \,. \tag{5.72}$$

Using (5.71) twice, we get:

$$\|\, f \,\|^2_{2,r} \leq \mathrm{e}^{2L} \left(1 + \frac{C_r}{c_r} \{g\}^2_{2,r}\right) \leq \mathrm{e}^{2\mathcal{C}_r K} \left(1 + \frac{C_r}{c_r} K^2\right).$$

Which proves the Lemma for $r \in (1/2, 1)$.

Second step : $r \in \mathbb{N}^*$

Let $r \in \mathbb{N}^*$. For $\ell \in W_r$, the r-th derivative of ℓ, say $\ell^{(r)}$, exists in $L^2(h)$ and:

$$\{\ell\}_{2,r}^2 = \|\,\ell^{(r)}\,\|_{L^2(h)}^2 \quad \text{as well as} \quad \|\,\ell\,\|_{2,r}^2 = \|\,\ell\,\|_{L^2(h)}^2 + \|\,\ell^{(r)}\,\|_{L^2(h)}^2 .$$

According to (5.43), we also get that for all $p \in \mathbb{N}$ with $p < r$ we have $\|\,\ell^{(p)}\,\|_\infty \le \mathcal{C}_{r-p}\{\ell^{(r)}\}_{2,r} \le \mathcal{C}_1\{\ell^{(r)}\}_{2,r}$.

Set $L = \mathcal{C}_r K$. Let $f = \mathrm{e}^g$ with $\|\,g\,\|_{2,r} \le K$. We have $\|\,g^{(p)}\,\|_\infty \le \mathcal{C}_1 K$ for all integer $p < r$. According to Leibniz's rule, we get that $f^{(r)} = g^{(r)} f + P_r(g^{(1)}, \dots, g^{(r-1)})f$, where P_r is a polynomial function of maximal degree r such that:

$$\max_{x_1,\dots,x_{r-1} \in [-\mathcal{C}_1 K, \mathcal{C}_1 K]} |P_r(x_1, \dots, x_{r-1})| \le C_{r,1} K^r. \tag{5.73}$$

for some finite constant $C_{r,1}$. We deduce that:

$$\|\,f^{(r)}\,\|_{L^2(h)} \le \mathrm{e}^L \,\|\,g^{(r)}\,\|_{L^2(h)} + \mathrm{e}^L\, C_{r,1} K^r.$$

Then use that $\|\,f\,\|_{L^2(h)} \le \mathrm{e}^L$ to get the Lemma for $r \in \mathbb{N}^*$.

Third step : $r > 1$, $r \notin \mathbb{N}^*$

Let $r > 1$ such that $r \notin \mathbb{N}^*$. Set $p = \lfloor r \rfloor \in \mathbb{N}^*$ the integer part of r and $s = r - p \in (0,1)$. For $\ell \in W_r$, the p-th derivative of ℓ, say $\ell^{(p)}$, exists in $L^2(h)$ and:

$$\{\ell\}_{2,r}^2 = \{\ell^{(p)}\}_{2,s}^2 \quad \text{as well as} \quad \|\,\ell\,\|_{2,r}^2 = \|\,\ell\,\|_{L^2(h)}^2 + \{\ell^{(p)}\}_{2,s}^2. \tag{5.74}$$

Thanks to (5.71) (twice) and the triangle inequality, we have for all measurable function t:

$$c_s\{\ell t\}_{2,s}^2 \le I_s(\ell t) \le \|\,t\,\|_\infty^2\, I_s(\ell) + J_s(\ell, t) \le \|\,t\,\|_\infty^2\, C_s\{\ell\}_{2,s}^2 + J_s(\ell, t), \tag{5.75}$$

with

$$J_s(\ell, t) = \frac{1}{2\pi} \int_{[-\pi,\pi]^2} \ell(x)^2\, \frac{|t(x+y) - t(x)|^2}{|y|^{1+2s}}\, dxdy.$$

Let $K > 0$ and set $L = \mathcal{C}_r K$. Let $f = \mathrm{e}^g$ with $g \in W_r$ such that $\|\,g\,\|_{2,r} \le K$. Following the proof of Lemma 5.17, we first give an upper bound of $J_s(\ell, f)$ in this context under the only condition that $\ell \in L^2(h)$. Using that $|\,\mathrm{e}^x - \mathrm{e}^y\,| \le \mathrm{e}^L\,|x - y|$ for $x, y \in [-L, L]$, we deduce that:

$$\int_{-\pi}^{\pi} \frac{|f(x+y) - f(x)|^2}{|y|^{1+2s}}\, dy \le \mathrm{e}^{2L} \int_{-\pi}^{\pi} \frac{|g(x+y) - g(x)|^2}{|y|^{1+2s}}\, dy.$$

Since a.e. $g(x) = \sum_{k\in\mathbb{Z}} a_k\, \mathrm{e}^{ikx}$, we deduce that:

$$J_s(\ell, f) \le \frac{\mathrm{e}^{2L}}{2\pi} \int_{-\pi}^{\pi} dx\, \ell(x)^2 \sum_{k,j\in\mathbb{Z}} |a_k||a_j| \int_{-\pi}^{\pi} \frac{|(1 - \mathrm{e}^{iky})(1 - \mathrm{e}^{-ijy})|}{|y|^{1+2s}}\, dy.$$

Let $\varepsilon \in (0, 1/2)$ such that $s + \varepsilon \le 1$. Since $|1 - \mathrm{e}^{ix}\,| \le 2|x|^{s+\varepsilon}$ for all $x \in \mathbb{R}$, we deduce that:

$$\int_{-\pi}^{\pi} \frac{|(1 - \mathrm{e}^{iky})(1 - \mathrm{e}^{-ijy})|}{|y|^{1+2s}}\, dy \le C_{2,\varepsilon} |k|^{s+\varepsilon} |j|^{s+\varepsilon},$$

for some constant $C_{2,\varepsilon}$ depending only on ε. Using Cauchy-Schwarz inequality and the fact that $r - s - \varepsilon > 1/2$, we get:

$$\sum_{k\in\mathbb{Z}} |k|^{s+\varepsilon} |a_k| \le \mathcal{C}_{r-s-\varepsilon}\{g\}_{2,r}.$$

We deduce that:

$$J_s(\ell, f) \le \mathrm{e}^{2L}\, \|\,\ell\,\|_{L^2(h)}^2\, C_{2,\varepsilon} \mathcal{C}_{r-s-\varepsilon}^2 \{g\}_{2,r}^2. \tag{5.76}$$

According to Leibniz's rule, we get that $f^{(p)} = \ell f + g^{(p)} f$ with $\ell = P_p(g^{(1)}, \dots, g^{(p-1)})$. We get:

$$c_s\{\ell f\}_{2,s}^2 \le \|f\|_\infty^2 C_s\{\ell\}_{2,s}^2 + J_s(\ell, f) \le \mathrm{e}^{2L} C_s\{f\}_{2,s}^2 + \mathrm{e}^{2L} \|\ell\|_{L^2(h)}^2 C_{2,\varepsilon}\mathcal{C}_{r-s-\varepsilon}^2\{g\}_{2,r}^2, \quad (5.77)$$

where we used (5.75) for the first inequality and (5.76) for the latter. Then use (5.73) with r replaced by p to get that $\|\ell\|_{L^2(h)} \le \|\ell\|_\infty \le C_{p,1}K^p$. Notice also that:

$$\{f\}_{2,s}^2 \le \mathrm{e}^{2L} \frac{C_s}{c_s}\{g\}_{2,s}^2,$$

using (5.71) twice and (5.72) (with s instead of r). We deduce that $\{\ell f\}_{2,s}$ is bounded by a constant depending only on K, r and ε.

The upper bound of $\{g^{(p)} f\}_{2,s}^2$ is similar. Using (5.75) and (5.76), we get:

$$c_s\{g^{(p)} f\}_{2,s}^2 \le \|f\|_\infty^2 I_s(g^{(p)}) + J_s(g^{(p)}, f) \le \mathrm{e}^{2L} C_s\{g^{(p)}\}_{2,s}^2 + \mathrm{e}^{2L} \|g^{(p)}\|_{L^2(h)}^2 C_{2,\varepsilon}\mathcal{C}_{r-s-\varepsilon}^2\{g\}_{2,r}^2.$$

We deduce that $\{g^{(p)} f\}_{2,s}$, and thus $f^{(p)}$, is bounded by a constant depending only on K, r and ε. Then use (5.74) and that $\|f\|_{L^2(h)} \le \|f\|_\infty \le \mathrm{e}^L$ to get the Lemma for $r > 1$ and $r \notin \mathbb{N}$. This concludes the proof.

Chapter 6

Fast adaptive estimation of log-additive exponential models in Kullback-Leibler divergence

6.1 Introduction

In this paper, we estimate densities with product form on the simplex $\triangle = \{(x_1, \dots, x_d) \in \mathbb{R}^d, 0 \leq x_1 \leq \cdots \leq x_d \leq 1\}$ by a nonparametric approach given a sample of n independent observations $\mathbb{X}^n = (X^1, \dots, X^n)$. We restrict our attention to densities which can be written in the form, for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$:

$$f^0(x) = \exp\left(\sum_{i=1}^{d} \ell_i^0(x_i) - \mathrm{a}_0\right)\mathbf{1}_{\triangle}(x), \tag{6.1}$$

with ℓ_i^0 bounded, centered, measurable functions on $I = [0, 1]$ for all $1 \leq i \leq d$, and normalizing constant a_0. Densities of this form arise, in particular, as solutions for the maximum entropy problem for the distribution of order statistics with given marginals, or in the case of the random truncation model.

The first example is the random truncation model, which was first formulated in [170], and has various applications ranging from astronomy ([126]), economics ([98], [90]) to survival data analysis ([118], [106], [125]). For $d = 2$, let (Z_1, Z_2) be a pair of independent random variables on I such that Z_i has density function p_i for $i \in \{1, 2\}$. Let us suppose that we can only observe realizations of (Z_1, Z_2) if $Z_1 \leq Z_2$. Let $(\bar{Z}_1, \bar{Z}_2)$ denote a pair of random variables distributed as (Z_1, Z_2) conditionally on $Z_1 \leq Z_2$. Then the joint density function f^0 of $(\bar{Z}_1, \bar{Z}_2)$ is given by, for $x = (x_1, x_2) \in I^2$:

$$f^0(x) = \frac{1}{\alpha}\, p_1(x_1)p_2(x_2)\mathbf{1}_{\triangle}(x), \tag{6.2}$$

with $\alpha = \int_{I^2} p_1(x_1)p_2(x_2)\mathbf{1}_{\triangle}(x)\, dx$. Notice that f is of the form required in (6.1):

$$f(x) = \exp(\ell_1^0(x_1) + \ell_2^0(x_2) - \mathrm{a}_0)\mathbf{1}_{\triangle}(x),$$

with ℓ_i^0 defined as $\ell_i^0 = \log(p_i) - \int_I \log(p_i)$ for $i \in \{1, 2\}$. According to Corollary 3.40, f is the density of the maximum entropy distribution of order statistics with marginals $\mathbf{f}_1$ and $\mathbf{f}_2$ given by:

$$\mathbf{f}_1(x_1) = \frac{1}{\alpha}\, p_1(x_1) \int_{x_1}^{1} p_2(s)\, ds \quad \text{and} \quad \mathbf{f}_2(x_2) = \frac{1}{\alpha}\, p_2(x_2) \int_{0}^{x_2} p_1(s)\, ds.$$

More generally, in [36], the authors give a necessary and sufficient condition for the existence of a maximum entropy distribution of order statistics with fixed marginal cumulative distribution functions $\mathbf{F}_i$, $1 \leq i \leq d$. See [34] for motivations for this problem. Moreover, its explicit expression

is given as a function of the marginal distributions. Let us suppose, for the sake of simplicity, that all $\mathbf{F}_i$ are absolutely continuous with density function $\mathbf{f}_i$ supported on $I = [0,1]$, and that $\mathbf{F}_{i-1} > \mathbf{F}_i$ on $(0,1)$ for $2 \leq i \leq d$. Then the maximum entropy density $f_{\mathbf{F}}$, when it exists, is given by, for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$:

$$f_{\mathbf{F}}(x) = \mathbf{f}_1(x_1) \prod_{i=2}^{d} h_i(x_i) \exp\left(- \int_{x_{i-1}}^{x_i} h_i(s)\, ds \right) \mathbf{1}_{\triangle}(x),$$

with $h_i = \mathbf{f}_i / (\mathbf{F}_{i-1} - \mathbf{F}_i)$ for $2 \leq i \leq d$. This density is of the form required in (6.1) with ℓ_i^0 defined as:

$$\ell_1^0 = \log(\mathbf{f}_1) + K_2 \quad \text{and} \quad \ell_i^0 = \log(h_i) - K_i + K_{i+1} \quad \text{for} \quad 2 \leq i \leq d,$$

with K_i, $2 \leq i \leq d$ a primitive of h_i chosen such that ℓ_i^0 are centered, and $K_{d+1} = c$ a constant.

We present a log-additive exponential model specifically designed to estimate such densities. This exponential model is a multivariate version of the exponential series estimator considered in [12] in the univariate setting. Essentially, we approximate the functions ℓ_i^0 by a family of polynomials $(\varphi_{i,k}, k \in \mathbb{N})$, which are orthonormal for each $1 \leq i \leq d$ with respect to the i-th marginal of the Lebesgue measure on the support $\triangle$. The model takes the form, for $\theta = (\theta_{i,k}; 1 \leq i \leq d, 1 \leq k \leq m_i)$ and $x = (x_1, \ldots, x_d) \in \triangle$:

$$f_\theta = \exp\left(\sum_{i=1}^{d} \sum_{k=1}^{m_i} \theta_{i,k} \varphi_{i,k}(x_i) - \psi(\theta) \right),$$

with $\psi(\theta) = \log\left(\int_{\triangle} \exp\left(\sum_{i=1}^{d} \sum_{k=1}^{m_i} \theta_{i,k} \varphi_{i,k}(x_i) \right) dx \right)$. Even though the polynomials $(\varphi_{i,k},\ k \in \mathbb{N})$ are orthonormal for each $1 \leq i \leq d$, if we take $i \neq j$, the families $(\varphi_{i,k}, k \in \mathbb{N})$ and $(\varphi_{j,k}, k \in \mathbb{N})$ are not completely orthogonal with respect to the Lebesgue measure on $\triangle$. The exact definition and further properties of these polynomials can be found in the Appendix. We estimate the parameters of the model by $\hat{\theta} = (\hat{\theta}_{i,k}; 1 \leq i \leq d, 1 \leq k \leq m_i)$, obtained by solving the maximum likelihood equations:

$$\int_{\triangle} \varphi_{i,k}(x_i) f_{\hat{\theta}}(x)\, dx = \frac{1}{n} \sum_{j=1}^{n} \varphi_{i,k}(X_i^j) \qquad \text{for } 1 \leq i \leq d,\ 1 \leq k \leq m_i.$$

Approximation of log-densities by polynomials appears in [87] as an application of the maximum entropy principle, while [47] shows existence and consistency of the maximum likelihood estimation. We measure the quality of the estimator $f_{\hat{\theta}}$ of f^0 by the Kullback-Leibler divergence $D\left(f^0 \| f_{\hat{\theta}}\right)$ defined as:

$$D\left(f^0 \| f_{\hat{\theta}}\right) = \int_{\triangle} f^0 \log\left(f^0 / f_{\hat{\theta}} \right).$$

Convergence rates for nonparametric density estimators have been given by [94] for kernel density estimators, [12] and [177] for the exponential series estimators, [13] for histogram-based estimators, and [116] for wavelet-based log-density estimators. Here, we give results for the convergence rate in probability when the functions ℓ_i^0 belong to a Sobolev space with regularity $r_i > d$ for all $1 \leq i \leq d$. We show that if we take $m = m(n) = (m_1(n), \ldots, m_d(n))$ members of the families $(\varphi_{i,k}, k \in \mathbb{N})$, $1 \leq i \leq d$, and let m_i grow with n such that $(\sum_{i=1}^{d} m_i^{2d})(\sum_{i=1}^{d} m_i^{-2r_i})$ and $(\sum_{i=1}^{d} m_i)^{2d+1}/n$ tend to 0, then the maximum likelihood estimator $f_{\hat{\theta}_{m,n}}$ verifies:

$$D\left(f^0 \| f_{\hat{\theta}_{m,n}}\right) = O_{\mathbb{P}}\left(\sum_{i=1}^{d} \left(m_i^{-2r_i} + \frac{m_i}{n} \right) \right).$$

Notice that this is the sum of the same univariate convergence rates as in [12]. By choosing m_i proportional to $n^{1/(2r_i+1)}$, which gives the optimal convergence rate $O_{\mathbb{P}}(n^{-2r_i/(2r_i+1)})$ in the

univariate case as shown in [181], we achieve a convergence rate of $O_{\mathbb{P}}(n^{-2\min(r)/(2\min(r)+1)})$. Therefore by exploiting the special structure of the underlying density, and carefully choosing the basis functions, we managed to reduce the problem of estimating a d-dimensional density to d one-dimensional density estimation problems. We highlight the fact that this constitutes a significant gain over convergence rates of general nonparametric multivariate density estimation methods.

In most cases the smoothness parameters r_i, $1 \leq i \leq d$, are not available, therefore a method which adapts to the unknown smoothness is required to estimate the density with the best possible convergence rate. Adaptive methods for function estimation based on a random sample include Lepski's method, model selection, wavelet thresholding and aggregation of estimators.

Lepski's method, originating from [123], consists of constructing a grid of regularities, and choosing among the minimax estimators associated to each regularity the best estimator by an iterative procedure based on the available sample. This method was extensively applied for Gaussian white noise model, regression, and density estimation, see [30] and references therein. Adaptation via model selection with a complexity penalization criterion was considered by [23] and [10] for a large variety of models including wavelet-based density estimation. Loss in the Kullback-Leibler distance for model selection was studied in [179] and [40] for mixing strategies, and in [182] for the information complexity minimization strategy. More recently, bandwidth selection for multivariate kernel density estimation was addressed in [84] for L^s risk, $1 \leq s < \infty$, and [122] for L^∞ risk. Wavelet based adaptive density estimation with thresholding was considered in [111] and [65], where an upper bound for the rate of convergence was given for a collection of Besov-spaces. Linear and convex aggregate estimators appear in the more recent work [150] with an application to adaptive density estimation in expected L^2 risk, with sample splitting.

Here we extend the convex aggregation scheme for the estimation of the logarithm of the density proposed in [38] to achieve adaptability. We take the estimator $f_{\hat{\theta}_{m,n}}$ for different values of $m \in \mathcal{M}_n$, where $\mathcal{M}_n$ is a sequence of sets of parameter configurations with increasing cardinality. These estimators are not uniformly bounded as required in [38], but we show that they are uniformly bounded in probability and that it does not change the general result. The different values of m correspond to different values of the regularity parameters. The convex aggregate estimator f_λ takes the form:

$$f_\lambda = \exp\left(\sum_{m\in\mathcal{M}_n} \lambda_m \left(\sum_{i=1}^{d}\sum_{k=1}^{m_i} \theta_{i,k}\varphi_{i,k}(x_i)\right) - \psi_\lambda\right)\mathbf{1}_\triangle,$$

with $\lambda \in \Lambda^+ = \{\lambda = (\lambda_m, m \in \mathcal{M}_n), \lambda_m \geq 0 \text{ and } \sum_{m\in\mathcal{M}_n} \lambda_m = 1\}$ and normalizing constant ψ_λ given by:

$$\psi_\lambda = \log\left(\int_\triangle \exp\left(\sum_{m\in\mathcal{M}_n} \lambda_m \left(\sum_{i=1}^{d}\sum_{k=1}^{m_i} \theta_{i,k}\varphi_{i,k}(x_i)\right)\right) dx\right).$$

To apply the aggregation method, we split our sample $\mathbb{X}^n$ into two parts $\mathbb{X}_1^n$ and $\mathbb{X}_2^n$, with size proportional to n. We use the first part to create the estimators $f_{\hat{\theta}_{m,n}}$, then we use the second part to determine the optimal choice of the aggregation parameter $\hat{\lambda}_n^*$. We select $\hat{\lambda}_n^*$ by maximizing a penalized version of the log-likelihood function. We show that this method gives a sequence of estimators $f_{\hat{\lambda}_n^*}$, free of the smoothness parameters $r_1, \ldots, r_d$, which verifies:

$$D\left(f^0 \| f_{\hat{\lambda}_n^*}\right) = O_{\mathbb{P}}\left(n^{-\frac{2\min(r)}{2\min(r)+1}}\right).$$

The rest of the paper is organized as follows. In Section 6.2 we introduce the notation used in the rest of the paper. In Section 6.3, we describe the log-additive exponential model and the estimation procedure, then we show that the estimator converges to the true underlying density with a convergence rate that is the sum of the convergence rates for the same type of univariate model, see Theorem 1.27. We consider an adaptive method with convex aggregation of

the logarithms of the previous estimators to adapt to the unknown smoothness of the underlying density in Section 6.4, see Theorem 6.8. We assess the performance of the adaptive estimator via a simulation study in Section 6.5. The definition of the basis functions and their properties used during the proofs are given in Section 6.6. The detailed proofs of the results in Section 6.3 and 6.4 are contained in Sections 6.7, 6.8 and 6.9.

6.2 Notation

Let $I = [0,1]$, $d \geq 2$ and $\triangle = \{(x_1, \ldots, x_d) \in I^d, x_1 \leq x_2 \leq \ldots \leq x_d\}$ denote the simplex of I^d. For an arbitrary real-valued function h_i defined on I with $1 \leq i \leq d$, let $h_{[i]}$ be the function defined on $\triangle$ such that for $x = (x_1, \ldots, x_d) \in \triangle$:

$$h_{[i]}(x) = h_i(x_i)\mathbf{1}_{\triangle}(x). \tag{6.3}$$

Let q_i, $1 \leq i \leq d$ be the one-dimensional marginals of the Lebesgue measure on $\triangle$:

$$q_i(dt) = \frac{1}{(d-i)!(i-1)!}(1-t)^{d-i}t^{i-1}\,\mathbf{1}_I(t)\,dt. \tag{6.4}$$

If $h_i \in L^1(q_i)$, then we have: $\int_{\triangle} h_{[i]} = \int_I h_i q_i$.

For a measurable function f, let $\|f\|_\infty$ be the usual sup norm of f on its domain of definition. For f defined on $\triangle$, let $\|f\|_{L^2} = \sqrt{\int_{\triangle} f^2}$. For f defined on I, let $\|f\|_{L^2(q_i)} = \sqrt{\int_I f^2 q_i}$.

For a vector $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, let $\min(r)$ ($\max(r)$) denote the smallest (largest) component.

Let us denote the support of a probability density g by $\text{supp}\,(g) = \{x \in \mathbb{R}^d, g(x) > 0\}$. Let $\mathcal{P}(\triangle)$ denote the set of probability densities on $\triangle$. For $g, h \in \mathcal{P}(\triangle)$, the Kullback-Leibler distance $D(g\|h)$ is defined as:

$$D(g\|h) = \int_{\triangle} g \log(g/h).$$

Recall that $D(g\|h) \in [0, +\infty]$.

Definition 6.1. We say that a probability density $f^0 \in \mathcal{P}(\triangle)$ has a product form if there exist $(\ell_i^0, 1 \leq i \leq d)$ bounded measurable functions defined on I such that $\int_I \ell_i^0 q_i = 0$ for $1 \leq i \leq d$ and a.e. on $\triangle$:

$$f^0 = \exp\left(\ell^0 - \mathrm{a}_0\right)\mathbf{1}_{\triangle}, \tag{6.5}$$

with $\ell^0 = \sum_{i=1}^d \ell^0_{[i]}$ and $\mathrm{a}_0 = \log\left(\int_{\triangle} \exp(\ell^0)\right)$, that is $f^0(x) = \exp\left(\sum_{i=1}^d \ell_i^0(x_i) - \mathrm{a}_0\right)$ for a.e. $x = (x_1, \ldots, x_d) \in \triangle$.

Definition 6.1 implies that $\text{supp}\,(f^0) = \triangle$ and f^0 is bounded. Let $\mathbb{X}^n = (X^1, \ldots, X^n)$ denote an i.i.d. sample of size n from the density f^0.

For $1 \leq i \leq d$, let $(\varphi_{i,k}, k \in \mathbb{N})$ be the family of orthonormal polynomials on I with respect to the measure q_i; see Section 6.6 for a precise definition of those polynomials and some of their properties. Recall $\varphi_{[i],k}(x) = \varphi_{i,k}(x_i)$ for $x = (x_1, \ldots, x_d) \in \triangle$. Notice that $(\varphi_{[i],k}, 1 \leq i \leq d, k \in \mathbb{N})$ is a family of normal polynomials with respect to the Lebesgue measure on $\triangle$, but not orthogonal.

Let $m = (m_1, \ldots, m_d) \in (\mathbb{N}^*)^d$ and set $|m| = \sum_{i=1}^d m_i$. We define the $\mathbb{R}^{|m|}$-valued function $\varphi_m = (\varphi_{[i],k}; 1 \leq k \leq m_i, 1 \leq i \leq d)$ and the $\mathbb{R}^{m_i}$-valued functions $\varphi_{i,m} = (\varphi_{i,k}; 1 \leq k \leq m_i)$ for $1 \leq i \leq d$. For $\theta = (\theta_{i,k}; 1 \leq k \leq m_i, 1 \leq i \leq d)$ and $\theta' = (\theta'_{i,k}; 1 \leq k \leq m_i, 1 \leq i \leq d)$ elements of $\mathbb{R}^{|m|}$, we denote the scalar product:

$$\theta \cdot \theta' = \sum_{i=1}^{d} \sum_{k=1}^{m_i} \theta_{i,k}\theta'_{i,k}$$

and the norm $\|\theta\| = \sqrt{\theta \cdot \theta}$. We define the function $\theta \cdot \varphi_m$ as follows, for $x \in \triangle$:

$$(\theta \cdot \varphi_m)(x) = \theta \cdot \varphi_m(x).$$

For a positive sequence $(a_n)_{n\in\mathbb{N}}$, the notation $O_{\mathbb{P}}(a_n)$ of stochastic boundedness for a sequence of random variables $(Y_n, n \in \mathbb{N})$ means that for every $\varepsilon > 0$, there exists $C_\varepsilon > 0$ such that:

$$\mathbb{P}\left(|Y_n/a_n| > C_\varepsilon\right) < \varepsilon \quad \text{for all } n \in \mathbb{N}.$$

6.3 Additive exponential series model

In this Section, we study the problem of estimation of an unknown density f^0 with a product form on the set $\triangle$, as described in (6.5), given the sample $\mathbb{X}^n$ drawn from f^0. Our goal is to give an estimation method based on a sequence of regular exponential models, which suits the special characteristics of the target density f^0. Estimating such a density with standard multidimensional nonparametric techniques naturally suffer from the curse of dimensionality, resulting in slow convergence rates for high-dimensional problems. We show that by taking into consideration that f^0 has a product form, we can recover the one-dimensional convergence rate for the density estimation, allowing for fast convergence of the estimator even if d is large. The quality of the estimators is measured by the Kullback-Leibler distance, as it has strong connections to the maximum entropy framework of [36].

We propose to estimate f^0 using the following log-additive exponential model, for $m \in (\mathbb{N}^*)^d$:

$$f_\theta = \exp\left(\theta \cdot \varphi_m - \psi(\theta)\right) \mathbf{1}_\triangle, \tag{6.6}$$

with $\psi(\theta) = \log\left(\int_\triangle \exp\left(\theta \cdot \varphi_m\right)\right)$. This model is similar to the one introduced in [177], but there are two major differences. First, we have only kept the univariate terms in the multivariate exponential series estimator of [177] since the target probability density is the product of univariate functions. Second, we have restricted our model to $\triangle$ instead of the hyper-cube I^d, and we have chosen the basis functions $((\varphi_{i,k}, k \in \mathbb{N}), 1 \le i \le d)$ which are appropriate for this support.

Remark 6.2. In the genaral case, one has to be careful when considering a density f^0 with a product form and a support different from $\triangle$. Let f_i^0 denote the i-th marginal density function of f^0. If supp $(f_i^0) = A \subset \mathbb{R}$ for all $1 \le i \le d$, we can apply a strictly monotone mapping of A onto I to obtain a distribution with a product form supported on $\triangle$. When the supports of the marginals differ, there is no transformation that yields a random vector with a density as in Definition 6.1. A possible way to treat this case consists of constructing a family of basis functions which has similar properties with respect to supp (f^0) as the family $((\varphi_{i,k}, k \in \mathbb{N}), 1 \le i \le d)$ with respect to $\triangle$, which we discuss in detail in Section 6.6. Then we could define an exponential series model with this family of basis functions and support restricted to supp (f^0) to estimate f^0.

Let $m \in (\mathbb{N}^*)^d$. We define the following function on $\mathbb{R}^{|m|}$ taking values in $\mathbb{R}^{|m|}$ by:

$$A_m(\theta) = \int_\triangle \varphi_m f_\theta, \quad \theta \in \mathbb{R}^{|m|}. \tag{6.7}$$

According to Lemma 3 in [12], we have the following result on A_m.

Lemma 6.3. *The function A_m is one-to-one from $\mathbb{R}^{|m|}$ to $\Omega_m = A_m(\mathbb{R}^{|m|})$.*

We denote by $\Theta_m : \Omega_m \mapsto \mathbb{R}^{|m|}$ the inverse of A_m. The empirical mean of the sample $\mathbb{X}^n$ of size n is:

$$\hat{\mu}_{m,n} = \frac{1}{n}\sum_{j=1}^{n} \varphi_m(X^j). \tag{6.8}$$

In Section 6.8.2 we show that $\hat{\mu}_{m,n} \in \Omega_m$ a.s. when $n \ge 2$.

For $n \geq 2$, we define a.s. the maximum likelihood estimator $\hat{f}_{m,n} = f_{\hat{\theta}_{m,n}}$ of f^0 by choosing:

$$\hat{\theta}_{m,n} = \Theta_m(\hat{\mu}_{m,n}). \tag{6.9}$$

The loss between the estimator $\hat{f}_{m,n}$ and the true underlying density f^0 is measured by the Kullback-Leibler divergence $D\left(f^0 \| \hat{f}_{m,n}\right)$.

For $r \in \mathbb{N}^*$, let $W_r^2(q_i)$ denote the Sobolev space of functions in $L^2(q_i)$, such that the $(r-1)$-th derivative is absolutely continuous and the L^2 norm of the r-th derivative is finite:

$$W_r^2(q_i) = \left\{ h \in L^2(q_i); h^{(r-1)} \text{ is absolutely continuous and } h^{(r)} \in L^2(q_i) \right\}.$$

The main result is given by the following theorem whose proof is given in Section 6.8.3.

Theorem 6.4. *Let $f^0 \in \mathcal{P}(\triangle)$ be a probability density with a product form, see Definition 6.1. Assume the functions ℓ_i^0, defined in (6.5) belong to the Sobolev space $W_{r_i}^2(q_i)$, $r_i \in \mathbb{N}$ with $r_i > d$ for all $1 \leq i \leq d$. Let $(X^n, n \in \mathbb{N}^*)$ be i.i.d. random variables with density distribution f^0. We consider a sequence $(m(n) = (m_1(n), \ldots, m_d(n)), n \in \mathbb{N}^*)$ such that $\lim_{n\to\infty} m_i(n) = +\infty$ for all $1 \leq i \leq d$, and which satisfies:*

$$\lim_{n\to\infty} |m|^{2d} \left(\sum_{i=1}^{d} m_i^{-2r_i} \right) = 0, \tag{6.10}$$

$$\lim_{n\to\infty} \frac{|m|^{2d+1}}{n} = 0. \tag{6.11}$$

The Kullback-Leibler distance $D\left(f^0 \| \hat{f}_{m,n}\right)$ of the maximum likelihood estimator $\hat{f}_{m,n}$ defined by (6.9) to f^0 converges in probability to 0 with the convergence rate:

$$D\left(f^0 \| \hat{f}_{m,n}\right) = O_{\mathbb{P}}\left(\sum_{i=1}^{d} m_i^{-2r_i} + \frac{|m|}{n} \right). \tag{6.12}$$

Remark 6.5. Let us take $(m^\circ(n) = (m_1^\circ(n), \ldots, m_d^\circ(n)), n \in \mathbb{N}^*)$ with $m_i^\circ(n) = \lfloor n^{1/(2r_i+1)} \rfloor$. This choice constitutes a balance between the bias and the variance term. Then the conditions (6.10) and (6.11) are satisfied, and we obtain that :

$$D\left(f^0 \| \hat{f}_{m^\circ,n}\right) = O_{\mathbb{P}}\left(\sum_{i=1}^{d} n^{-2r_i/(2r_i+1)} \right) = O_{\mathbb{P}}\left(n^{-2\min(r)/(2\min(r)+1)} \right).$$

Thus the convergence rate corresponds to the least smooth ℓ_i^0. This rate can also be obtained with a choice where all m_i are the same. Namely, with $(m^*(n) = (v^*(n), \ldots, v^*(n)), n \in \mathbb{N}^*)$ and $v^*(n) = \lfloor n^{1/(2\min(r)+1)} \rfloor$.

For $r = (r_1, \ldots, r_d) \in (\mathbb{N}^*)^d$, $r_i > d$ for $1 \leq i \leq d$, and a constant $\kappa > 0$, let :

$$\mathcal{K}_r(\kappa) = \left\{ f^0 = \exp\left(\sum_{i=1}^{d} \ell_{[i]}^0 - \mathrm{a}_0 \right) \in \mathcal{P}(\triangle); \| \ell_i^0 \|_\infty \leq \kappa, \| (\ell_i^0)^{(r_i)} \|_{L^2(q_i)} \leq \kappa \right\}. \tag{6.13}$$

The constants $\mathfrak{A}_1$ and $\mathfrak{A}_2$, appearing in the upper bounds during the proof of Theorem 6.4 (more precisely in Propositions 6.35 and 6.37), are uniformly bounded on $\mathcal{K}_r(\kappa)$, thanks to Corollary 6.24 and $\| \log(f^0) \|_\infty \leq 2d\kappa + |\log(d!)|$, which is due to (6.43). This yields the following corollary for the uniform convergence in probability on the set $\mathcal{K}_r(\kappa)$ of densities:

Corollary 6.6. *Under the assumptions of Theorem 6.4, we get the following result:*

$$\lim_{K\to\infty} \limsup_{n\to\infty} \sup_{f^0 \in \mathcal{K}_r(\kappa)} \mathbb{P}\left(D\left(f^0 \| \hat{f}_{m,n}\right) \geq \left(\sum_{i=1}^{d} m_i^{-2r_i} + \frac{|m|}{n} \right) K \right) = 0.$$

Remark 6.7. Since we let r_i vary for each $1 \leq i \leq d$, our class of densities $\mathcal{K}_r(\kappa)$ has an anisotropic feature. Estimation of anisotropic multivariate functions for L^s risk, $1 \leq s \leq \infty$, was considered in multiple papers. For a Gaussian white noise model, [112] obtains minimax convergence rates on anisotropic Besov classes for L^s risk, $1 \leq s < \infty$,while [19] gives the minimax rate of convergence on anisotropic Hölder classes for the L^∞ risk. For kernel density estimation, results on the minimax convergence rate for anisotropic Nikol'skii classes for L^s risk, $1 \leq s < \infty$, can be found in [84]. These papers conclude in general, that if the considered class has smoothness parameters $\tilde{r}_i$ for the i-th coordinate, $1 \leq i \leq d$, then the optimal convergence rate becomes $n^{-2\tilde{R}/(2\tilde{R}+1)}$ (multiplied with a logarithmic factor for L^∞ risk), with $\tilde{R}$ defined by the equation $1/\tilde{R} = \sum_{i=1}^d 1/\tilde{r}_i$. Since $\tilde{R} < \tilde{r}_i$ for all $1 \leq i \leq d$, the convergence rate $n^{-2\min(r)/(2\min(r)+1)}$ is strictly better than the convergence rate for these anisotropic classes. In the isotropic case, when $r_i = r$ for all $1 \leq i \leq d$, the minimax convergence rate specializes to $n^{-2r/(2r+d)}$ (which was obtained in [177] as an upper bound). This rate decreases exponentially when the dimension d increases. However, by exploiting the multiplicative structure of the model, we managed to obtain the univariate convergence rate $n^{-2r/(2r+1)}$, which is minimax optimal, see [181].

6.4 Adaptive estimation

Notice that the choice of the optimal series of estimators $\hat{f}_{m^*,n}$ with m^* defined in Remark 6.5 requires the knowledge of $\min(r)$ at least. When this knowledge is not available, we propose an adaptive method based on the proposed estimators in Section 6.3, which can mimic asymptotically the behaviour of the optimal choice. Let us introduce some notation first. We separate the sample $\mathbb{X}^n$ into two parts $\mathbb{X}_1^n$ and $\mathbb{X}_2^n$ of size $n_1 = \lfloor C_e n \rfloor$ and $n_2 = n - \lfloor C_e n \rfloor$ respectively, with some constant $C_e \in (0,1)$. The first part of the sample will be used to create our estimators, and the second half will be used in the aggregation procedure. Let $(N_n, n \in \mathbb{N}^*)$ be a sequence of non-decreasing positive integers depending on n such that $\lim_{n\to\infty} N_n = +\infty$. Let us denote:

$$\mathcal{N}_n = \left\{ \lfloor n^{1/(2(d+j)+1)} \rfloor, 1 \leq j \leq N_n \right\} \quad \text{and} \quad \mathcal{M}_n = \left\{ m = (v, \dots, v) \in \mathbb{R}^d, v \in \mathcal{N}_n \right\}. \tag{6.14}$$

For $m \in \mathcal{M}_n$ let $\hat{f}_{m,n}$ be the maximum likelihood estimator for the log-additive exponential model based on the first half of the sample, namely:

$$\hat{f}_{m,n} = \exp\left(\hat{\theta}_{m,n} \cdot \varphi_m - \psi(\hat{\theta}_{m,n}) \right) \mathbf{1}_\triangle,$$

with $\hat{\theta}_{m,n}$ given by (6.9) using the sample $\mathbb{X}_1^n$ (replacing n with n_1 in the definition (6.8) of $\hat{\mu}_{m,n}$). Let :

$$\mathcal{F}_n = \{\hat{f}_{m,n}, m \in \mathcal{M}_n\}$$

denote the set of different estimators obtained by this procedure. Notice that Card $(\mathcal{F}_n) \leq$ Card $(\mathcal{M}_n) \leq N_n$. Recall that by Remark 6.5, we have that for $r = (r_1, \dots, r_d)$ with $r_i > d$ and $n \geq \bar{n}$, where $\bar{n}$ is given by:

$$\bar{n} = \min\{n \in \mathbb{N}, N_n \geq \min(r) - d + 1\}, \tag{6.15}$$

the sequence of estimators $\hat{f}_{m^*,n}$, with $m^* = m^*(n) = (v^*, \dots, v^*) \in \mathcal{M}_n$ given by $v^* = \lfloor n^{1/(2\min(r)+1)} \rfloor$, achieves the optimal convergence rate $O_{\mathbb{P}}(n^{-2\min(r)/(2\min(r)+1)})$. By letting N_n go to infinity, we ensure that for every combination of regularity parameters $r = (r_1, \dots, r_d)$ with $r_i > d$, the sequence of optimal estimators $\hat{f}_{m^*,n}$ is included in the sets $\mathcal{F}_n$ for n large enough.

We use the second part of the sample $\mathbb{X}_2^n$ to create an aggregate estimator based on $\mathcal{F}_n$, which asymptotically mimics the performance of the optimal sequence $\hat{f}_{m^*,n}$. We will write $\hat{\ell}_{m,n} = \hat{\theta}_{m,n} \cdot \varphi_m$ to ease notation. We define the convex combination $\hat{\ell}_\lambda$ of the functions $\hat{\ell}_{m,n}$, $m \in \mathcal{M}_n$:

$$\hat{\ell}_\lambda = \sum_{m \in \mathcal{M}_n} \lambda_m \hat{\ell}_{m,n},$$

with aggregation weights $\lambda \in \Lambda^+ = \{\lambda = (\lambda_m, m \in \mathcal{M}_n) \in \mathbb{R}^{\mathcal{M}_n}, \lambda_m \geq 0 \text{ and } \sum_{m\in\mathcal{M}_n} \lambda_m = 1\}$. For such a convex combination, we define the probability density function f_λ as:

$$f_\lambda = \exp(\hat{\ell}_\lambda - \psi_\lambda)\mathbf{1}_\triangle, \tag{6.16}$$

with $\psi_\lambda = \log\left(\int_\triangle \exp(\hat{\ell}_\lambda)\right)$. We apply the convex aggregation method for log-densities developed in [38] to get an aggregate estimator which achieves adaptability. Notice that the reference probability measure in this paper corresponds to $d!\mathbf{1}_\triangle(x)dx$. This implies that ψ_λ here differs from the ψ_λ of [38] by the constant $\log(d!)$, but this does not affect the calculations. The aggregation weights are chosen by maximizing the penalized maximum likelihood criterion H_n defined as:

$$H_n(\lambda) = \frac{1}{n_2} \sum_{X^j \in \mathbb{X}_2^n} \hat{\ell}_\lambda(X^j) - \psi_\lambda - \frac{1}{2}\,\mathrm{pen}(\lambda), \tag{6.17}$$

with the penalizing function $\mathrm{pen}(\lambda) = \sum_{m\in\mathcal{M}_n} \lambda_m\, D\left(f_\lambda \| \hat{f}_{m,n}\right)$. The convex aggregate estimator $f_{\hat{\lambda}_n^*}$ is obtained by setting:

$$\hat{\lambda}_n^* = \underset{\lambda\in\Lambda^+}{\mathrm{argmax}}\ H_n(\lambda). \tag{6.18}$$

The main result of this section is given by the next theorem which asserts that if we choose $N_n = o(\log(n))$ such that $\lim_{n\to\infty} N_n = +\infty$, the series of convex aggregate estimators $f_{\hat{\lambda}_n^*}$ converge to f^0 with the optimal convergence rate, i.e. as if the smoothness was known.

Theorem 6.8. *Let $f^0 \in \mathcal{P}(\triangle)$ be a probability density with a product form given by (6.5). Assume the functions ℓ_i^0 belongs to the Sobolev space $W_{r_i}^2(q_i)$, $r_i \in \mathbb{N}$ with $r_i > d$ for all $1 \leq i \leq d$. Let $(X^n, n \in \mathbb{N}^*)$ be i.i.d. random variables with density f^0. Let $N_n = o(\log(n))$ such that $\lim_{n\to\infty} N_n = +\infty$. The convex aggregate estimator $f_{\hat{\lambda}_n^*}$ defined by (6.16) with $\hat{\lambda}_n^*$ given by (6.18) converges to f^0 in probability with the convergence rate:*

$$D\left(f^0 \| f_{\hat{\lambda}_n^*}\right) = O_{\mathbb{P}}\left(n^{-\frac{2\min(r)}{2\min(r)+1}}\right). \tag{6.19}$$

The proof of this theorem is provided in Section 6.9. Similarly to Corollary 6.6, we have uniform convergence over sets of densities with increasing regularity. Recall the definition (6.13) of the set $\mathcal{K}_r(\kappa)$. Let $\mathcal{R}_n = \{j, d+1 \leq j \leq R_n\}$, where R_n satisfies the three inequalities:

$$R_n \leq N_n + d, \tag{6.20}$$

$$R_n \leq \left\lfloor n^{\frac{1}{2(d+N_n)+1}} \right\rfloor, \tag{6.21}$$

$$R_n \leq \frac{\log(n)}{2\log(\log(N_n))} - \frac{1}{2}\,. \tag{6.22}$$

Corollary 6.9. *Under the assumptions of Theorem 6.8, we get the following result:*

$$\lim_{K\to\infty} \limsup_{n\to\infty} \sup_{r\in(\mathcal{R}_n)^d} \sup_{f^0\in\mathcal{K}_r(\kappa)} \mathbb{P}\left(D\left(f^0 \| f_{\hat{\lambda}_n^*}\right) \geq \left(n^{-\frac{2\min(r)}{2\min(r)+1}}\right) K\right) = 0.$$

Remark 6.10. For example when $N_n = \log(n)/(2\log(\log(n)))$, then (6.20), (6.21) and (6.22) are satisfied with $R_n = N_n$ for n large enough.

6.5 Simulation study : random truncation model

In this section we present the results of Monte Carlo simulation studies on the performance of the maximum likelihood estimator of the log-additive exponential model. We take the example of the random truncation model introduced in Section 6.1 with $d = 2$, which is used in many applications. This model naturally satisfies our model assumptions.

Let $Z = (Z_1, Z_2)$ be a pair of independent random variable with density functions q_1, q_2 respectively such that $\triangle \subset \text{supp}\,(q)$, where $q(x_1, x_2) = q_1(x_1)q_2(x_2)$ is the joint density function of Z. Suppose that we only observe pairs (Z_1, Z_2) if $0 \leq Z_1 \leq Z_2 \leq 1$. Then the joint density function f of the observable pairs is given by, for $x = (x_1, x_2) \in \mathbb{R}^2$:

$$f(x) = \frac{q_1(x_1)q_2(x_2)}{\int_{\triangle} q(y)\,dy}\mathbf{1}_{\triangle}(x).$$

This corresponds to the form (6.2) with p_1, p_2 given by:

$$p_1 = \frac{q_1\mathbf{1}_I}{\int_I q_1} \quad \text{and} \quad p_2 = \frac{q_2\mathbf{1}_I}{\int_I q_2}.$$

We will choose the densities q_1, q_2 from the following distributions:

- Normal(μ, σ^2) with $\mu \in \mathbb{R}, \sigma > 0$:

$$f_{\mu,\sigma^2}(t) = \frac{1}{\sqrt{2\pi\sigma^2}}\,\mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}},$$

- NormalMix$(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, w)$ with $w \in (0, 1)$:

$$f(t) = w f_{\mu_1,\sigma_1^2}(t) + (1-w) f_{\mu_2,\sigma_2^2}(t),$$

- Beta(α, β, a, b) with $0 < \alpha < \beta$, $a < 0$, $b > 1$:

$$f(t) = \frac{(t-a)^{\alpha-1}(b-t)^{\beta-\alpha-1}}{(b-a)^{\beta-1}B(\alpha, \beta-\alpha)}\mathbf{1}_{(a,b)}(t),$$

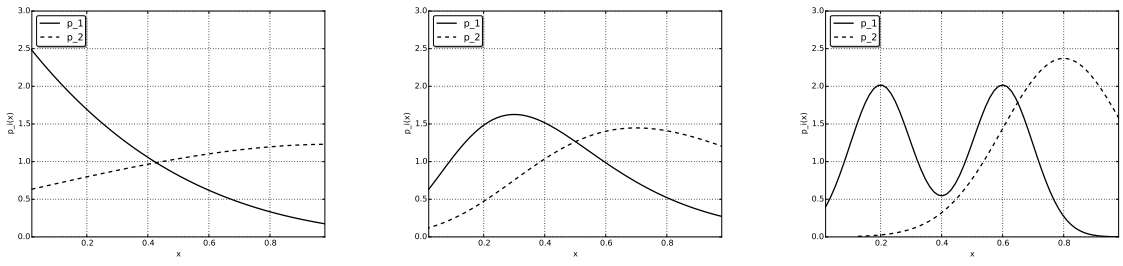
- Gumbel(α, β) with $\alpha > 0$, $\beta \in \mathbb{R}$:

$$f(t) = \alpha\,\mathrm{e}^{-\alpha(t-\beta)-\mathrm{e}^{-\alpha(t-\beta)}}.$$

The exact choices for densities q_1, q_2 are given in Table 6.1. Figure 6.1 shows the resulting density functions p_1 and p_2 for each case.

Model	q_1	q_2
Beta	Beta$(1, 6, -1, 2)$	Beta$(3, 5, -1, 2)$
Gumbel	Gumbel$(4, 0.3)$	Gumbel$(2.4, 0.7)$
Normal mix	NormalMix$(0.2, 0.1, 0.6, 0.1, 0.5)$	Normal$(0.8, 0.2)$

Table 6.1 – Distributions for the left-truncated model used in the simulation study.



(a) Beta (b) Gumbel (c) Normal mix

Figure 6.1 – Density functions p_1, p_2 of the left-truncated models used in the simulation study.

To calculate the parameters $\hat{\theta}_{m,n}$, we recall that $\hat{\theta}_{m,n}$ is the solution of the equation (6.9), therefore can be also characterized as:

$$\hat{\theta}_{m,n} = \text{argmax}_{\,\theta\in\mathbb{R}^{|m|}}\theta \cdot \hat{\mu}_{m,n} - \psi(\theta), \tag{6.23}$$

with $\hat{\mu}_{m,n}$ defined by (6.8), see Lemma 6.28 . We use a numerical optimisation method to solve (6.23) and obtain the parameters $\hat{\theta}_{m,n}$. We estimate our model with $m_1 = m_2 = \bar{m}$, and $\bar{m} = 1, 2, 3, 4$. We compute the final estimator based on the convex aggregation method proposed in Section 6.4. We ran 100 estimations with increasing sample sizes $n \in \{200, 500, 1000\}$, and we calculated the average Kullback-Leibler distance as well as the L^2 distance between f^0 and its estimator. We used $C_e = 80\%$ of the sample to calculate the initial estimators, and the remaining 20% to perform the aggregation. The distances were calculated by numerical integration. We compare the results with a truncated kernel density estimator with Gaussian kernel functions and bandwidth selection based on Scott's rule. The results are summarized in Table 6.2 and Table 6.3.

Table 6.2 – Average Kullback-Leibler distances for the log-additive exponential series estimator (LAESE) and the truncated kernel estimator (Kernel) based on 100 samples of size n. Variances provided in parenthesis.

KL distances	n=200		n=500		n=1000	
	LAESE	Kernel	LAESE	Kernel	LAESE	Kernel
Beta	0.0137 (8.94E-05)	0.0524 (1.73E-04)	0.0048 (9.51E-06)	0.0395 (4.61E-05)	0.0028 (3.50E-06)	0.0339 (2.14E-05)
Gumbel	0.0204 (1.48E-04)	0.0249 (8.03E-05)	0.0089 (2.88E-05)	0.0180 (2.07E-05)	0.0050 (6.70E-06)	0.0154 (1.03E-05)
Normal mix	0.0545 (4.51E-04)	0.0774 (7.29E-05)	0.0337 (1.88E-04)	0.0559 (2.95E-05)	0.0259 (2.50E-05)	0.0433 (1.52E-05)

Table 6.3 – Average L^2 distances for the log-additive exponential series estimator (LAESE) and the truncated kernel estimator (Kernel) based on 100 samples of size n. Variances provided in parenthesis.

$\mathbb{L}^2$ distances	n=200		n=500		n=1000	
	LAESE	Kernel	LAESE	Kernel	LAESE	Kernel
Beta	0.0536 (1.42E-03)	0.2107 (2.60E-03)	0.0200 (2.27E-04)	0.1660 (8.04E-04)	0.0120 (7.45E-05)	0.1429 (3.52E-04)
Gumbel	0.0683 (1.95E-03)	0.0856 (9.94E-04)	0.0297 (3.61E-04)	0.0621 (2.49E-04)	0.0166 (8.74E-05)	0.0522 (1.19E-04)
Normal mix	0.2314 (1.17E-02)	0.3534 (1.43E-03)	0.1489 (5.53E-03)	0.2545 (6.95E-04)	0.1112 (9.25E-04)	0.1952 (3.83E-04)

We can conclude that the log-additive exponential series estimator outperforms the kernel density estimator both with respect to the Kullback-Leibler distance and the L^2 distance. As expected, the performance of both methods increases with the sample size. The boxplot of the 100 values of the Kullback-Leibler and L^2 distance for the different sample sizes can be found in Figures 6.2, 6.4 and 6.6. Figures 6.3, 6.5 and 6.7 illustrate the different estimators compared to the true joint density function for the three cases obtained with a sample size of 1000. We can observe that the log.additive exponential model leads to a smooth estimator compared to the kernel method.

Remark 6.11. The log-additive exponential model encompasses a lot of popular choices for the marginals q_1, q_2. For example, the exponential distribution is included in the model for $m_i = 1$, and the normal distribution is included for $m_i = 2$. Thus we expect that if we choose exponential or normal distributions for q_1, q_2, we obtain even better results for the log-additive

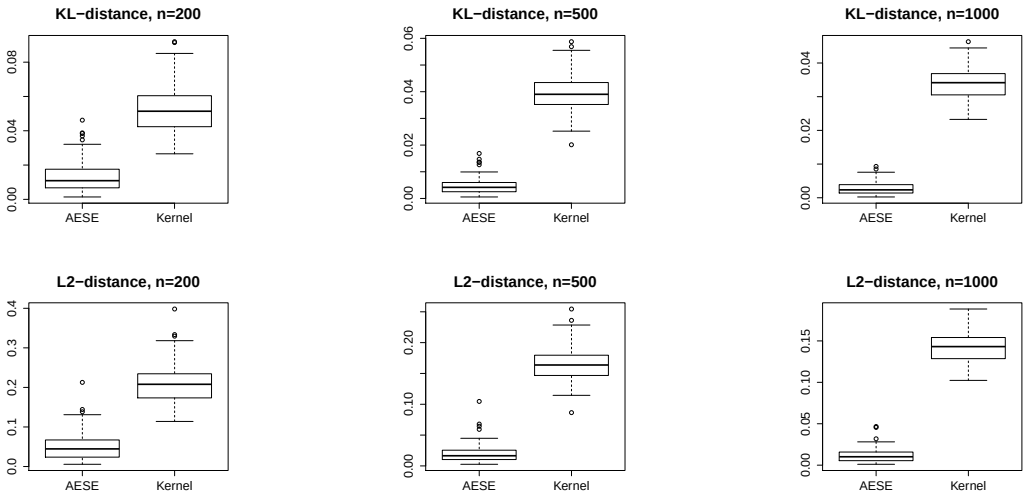
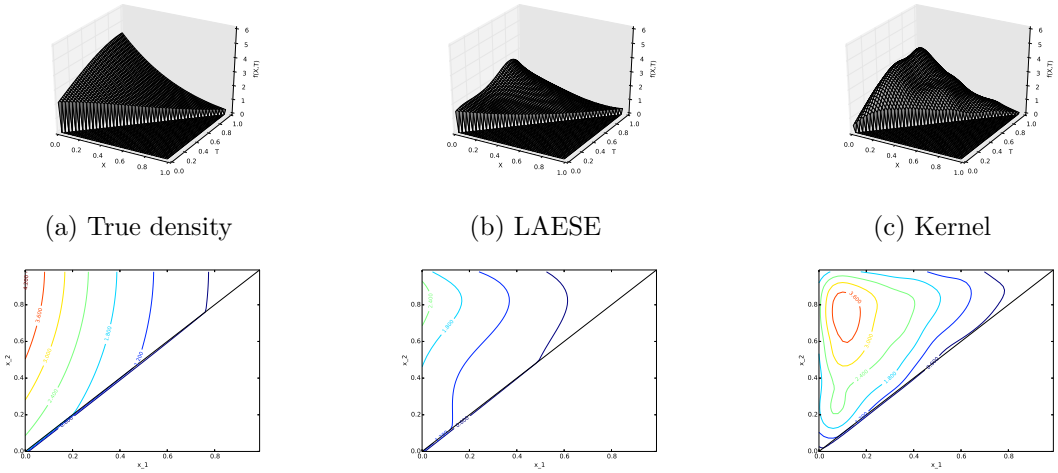


Figure 6.2 – Boxplot of the Kullback-Leibler and L^2 distances for the additive exponential series estimator (AESE) and the truncated kernel estimators with Beta marginals.



(a) True density (b) LAESE (c) Kernel

(d) True density (e) LAESE (f) Kernel

Figure 6.3 – Joint density functions of the true density and its estimators with Beta marginals.

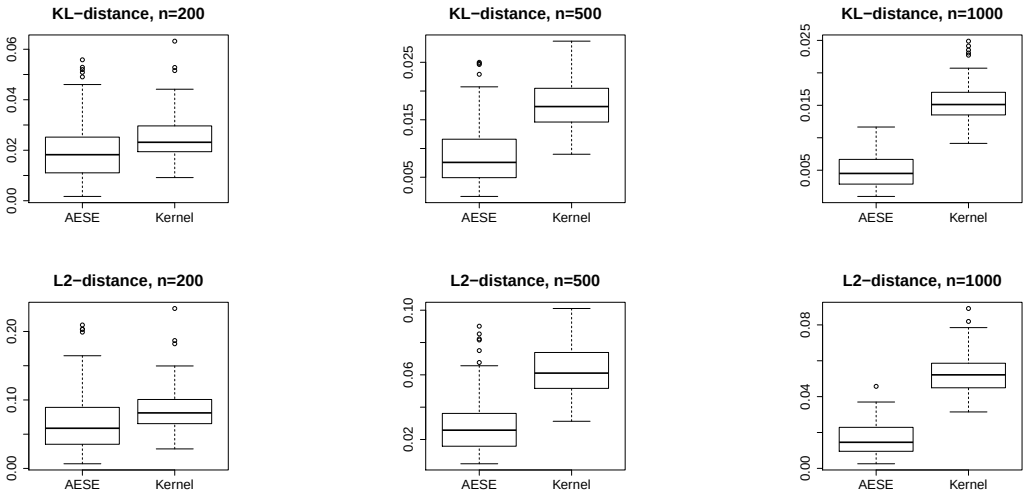
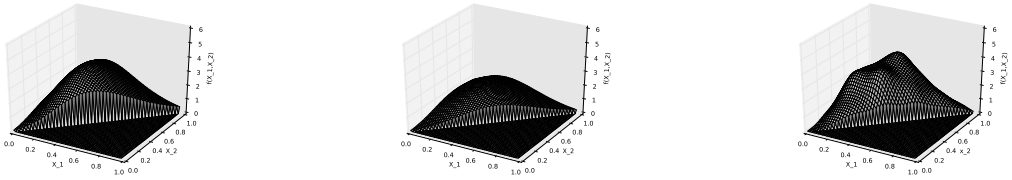
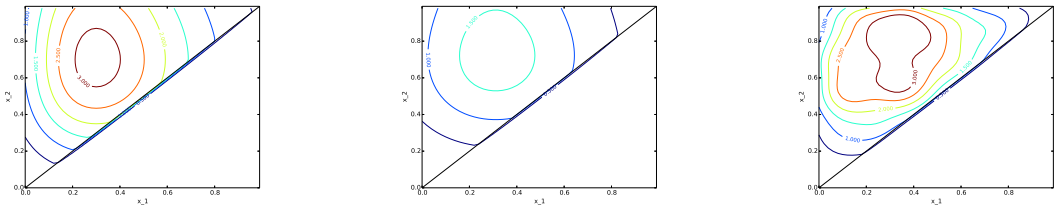


Figure 6.4 – Boxplot of the Kullback-Leibler and L^2 distances for the additive exponential series estimator (AESE) and the truncated kernel estimators with Gumbel marginals.



(a) True density (b) LAESE (c) Kernel



(d) True density (e) LAESE (f) Kernel

Figure 6.5 – Joint density functions of the true density and its estimators with Gumbel marginals.

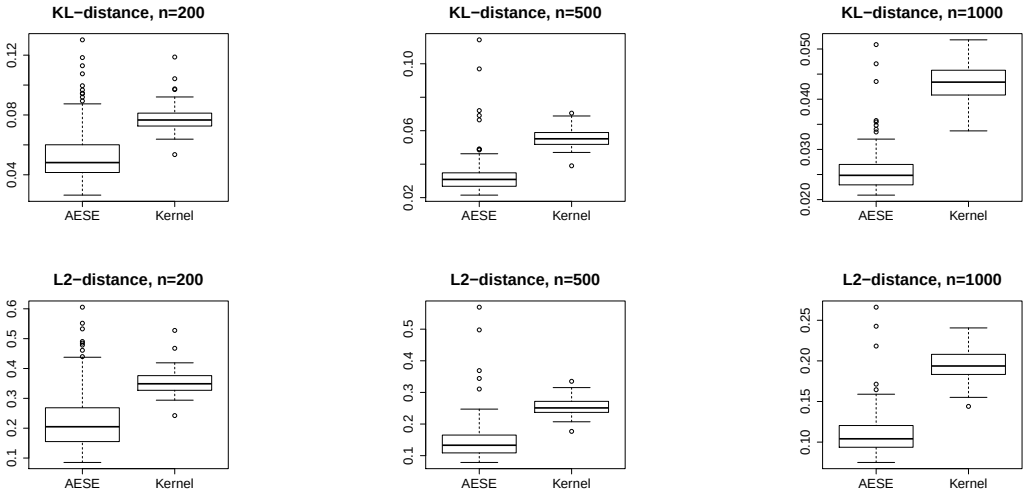
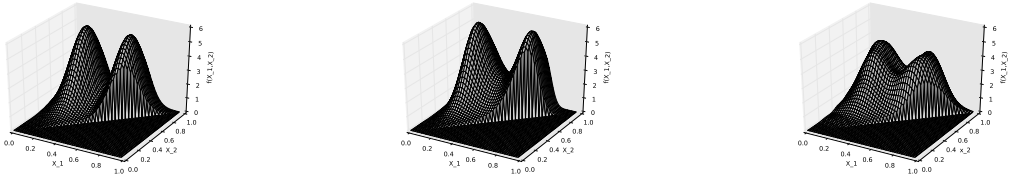
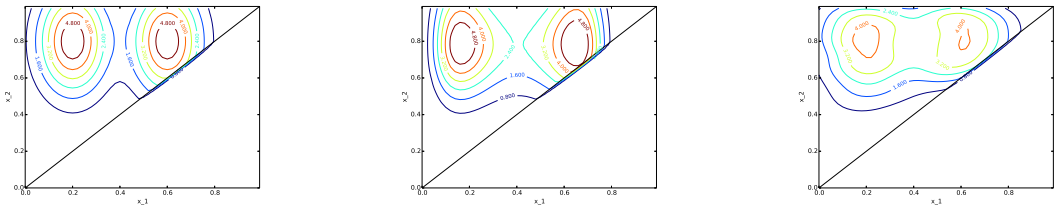


Figure 6.6 – Boxplot of the Kullback-Leibler and L^2 distances for the additive exponential series estimator (AESE) and the truncated kernel estimators with Normal mix marginals.



(a) True density (b) LAESE (c) Kernel



(d) True density (e) LAESE (f) Kernel

Figure 6.7 – Joint density functions of the true density and its estimators with Normal mix marginals.

exponential series estimator, which was confirmed by the numerical experiments (not included here for brevity).

6.6 Appendix: Orthonormal series of polynomials

6.6.1 Jacobi polynomials

The following results can be found in [2] p. 774. The Jacobi polynomials $(P_k^{(\alpha,\beta)}, k \in \mathbb{N})$ for $\alpha, \beta \in (-1, +\infty)$ are series of orthogonal polynomials with respect to the measure $w_{\alpha,\beta}(t)\mathbf{1}_{[-1,1]}(t)\,dt$, with $w_{\alpha,\beta}(t) = (1-t)^\alpha(1+t)^\beta$ for $t \in [-1,1]$. They are given by Rodrigues' formula, for $t \in [-1,1]$, $k \in \mathbb{N}$:

$$P_k^{(\alpha,\beta)}(t) = \frac{(-1)^k}{2^k k! w_{\alpha,\beta}(t)} \frac{d^k}{dt^k}\left[w_{\alpha,\beta}(t)(1-t^2)^k\right].$$

The normalizing constants are given by:

$$\int_{-1}^{1} P_k^{(\alpha,\beta)}(t) P_\ell^{(\alpha,\beta)}(t) w_{\alpha,\beta}(t)\,dt = \mathbf{1}_{\{k=\ell\}} \frac{2^{\alpha+\beta+1}}{2k+\alpha+\beta+1} \frac{\Gamma(k+\alpha+1)\Gamma(k+\beta+1)}{\Gamma(k+\alpha+\beta+1)k!}. \tag{6.24}$$

In what follows, we will be interested in Jacobi polynomials with $\alpha = d-i$ and $\beta = i-1$, which are orthogonal to the weight function $w_{d-i,i-1}(t) = \mathbf{1}_{[-1,1]}(t)(1-t)^{d-i}(1+t)^{i-1}$. The leading coefficient of $P_k^{(d-i,i-1)}$ is:

$$\omega'_{i,k} = \frac{(2k+d-1)!}{2^k k!(k+d-1)!}. \tag{6.25}$$

Let $r \in \mathbb{N}^*$. Recall that $P_k^{(\alpha,\beta)}$ has degree k. The derivatives of the Jacobi polynomials $P_k^{(d-i,i-1)}$, $r \le k$, verify, for $t \in I$ (see Proposition 1.4.15 of [67]):

$$\frac{d^r}{dt^r} P_k^{(d-i,i-1)}(t) = \frac{(k+d-1+r)!}{2^r(k+d-1)!} P_{k-r}^{(d-i+r,i-1+r)}(t). \tag{6.26}$$

We also have:

$$\sup_{t\in[-1,1]} \left|P_k^{(d-i,i-1)}(t)\right| = \max\left(\frac{(k+d-i)!}{k!(d-i)!}, \frac{(k+i-1)!}{k!(i-1)!}\right). \tag{6.27}$$

6.6.2 Definition of the basis functions

Based on the Jacobi polynomials, we define a shifted version, normalized and adapted to the interval $I = [0,1]$.

Definition 6.12. For $1 \le i \le d$, $k \in \mathbb{N}$, we define for $t \in I$:

$$\varphi_{i,k}(t) = \rho_{i,k}\sqrt{(d-i)!(i-1)!}\, P_k^{(d-i,i-1)}(2t-1),$$

with

$$\rho_{i,k} = \sqrt{(2k+d)k!(k+d-1)!/((k+d-i)!(k+i-1)!)}. \tag{6.28}$$

Recall the definition (6.4) of the marginals q_i of the Lebesgue measure on the simplex. According to the following Lemma, the polynomials $(\varphi_{i,k}, k \in \mathbb{N})$ form an orthonormal basis of $L^2(q_i)$ for all $1 \le i \le d$. Notice that $\varphi_{i,k}$ has degree k.

Lemma 6.13. *For $1 \le i \le d$, $k, \ell \in \mathbb{N}$, we have:*

$$\int_I \varphi_{i,k}\varphi_{i,\ell}\, q_i = \mathbf{1}_{\{k=\ell\}}.$$

Proof. We have, for $k, \ell \in \mathbb{N}$:

$$\begin{aligned}\int_I \varphi_{i,k}\varphi_{i,\ell}\, q_i &= \rho_{i,k}\rho_{i,\ell}\int_0^1 P_k^{(d-i,i-1)}(2t-1)P_\ell^{(d-i,i-1)}(2t-1)(1-t)^{d-i}t^{i-1}\,dt \\ &= \frac{\rho_{i,k}\rho_{i,\ell}}{2^d}\int_{-1}^1 P_k^{(d-i,i-1)}(s)P_\ell^{(d-i,i-1)}(s)w_{d-i,i-1}(s)\,ds \\ &= \mathbf{1}_{\{k=\ell\}},\end{aligned}$$

where we used (6.24) for the last equality. □

6.6.3 Mixed scalar products

Recall notation (6.3), so that $\varphi_{[i],k}(x) = \varphi_{i,k}(x_i)$ for $x = (x_1, \ldots, x_d) \in \triangle$. Notice that $(\varphi_{[i],k}, k \in \mathbb{N})$ is a family of orthonormal polynomials with respect to the Lebesgue measure on $\triangle$, for all $1 \le i \le d$.

We give the mixed scalar products of $(\varphi_{[i],k}, k \in \mathbb{N})$ and $(\varphi_{[j],\ell}, \ell \in \mathbb{N})$, $1 \le i < j \le d$ with respect to the Lebesgue measure on the simplex $\triangle$.

Lemma 6.14. *For $1 \le i < j \le d$ and $k, \ell \in \mathbb{N}$, we have:*

$$\int_\triangle \varphi_{[i],k}\,\varphi_{[j],\ell} = \mathbf{1}_{\{k=\ell\}}\sqrt{\frac{(j-1)!(d-i)!}{(i-1)!(d-j)!}}\sqrt{\frac{(k+d-j)!(k+i-1)!}{(k+d-i)!(k+j-1)!}}.$$

We also have $0 \le \int_\triangle \varphi_{[i],k}\,\varphi_{[j],\ell} \le 1$ for all $k, \ell \in \mathbb{N}$.

Proof. We have:

$$\begin{aligned}\int_\triangle \varphi_{[i],k}\,\varphi_{[j],\ell} &= \int_0^1 \left(\int_0^{x_j} \frac{x_i^{i-1}}{(i-1)!}\frac{(x_j-x_i)^{j-i-1}}{(j-i-1)!}\varphi_{i,k}(x_i)\,dx_i\right)\varphi_{j,\ell}(x_j)\frac{(1-x_j)^{d-j}}{(d-j)!}\,dx_j \\ &= \int_I r_k\varphi_{j,\ell}\, q_j,\end{aligned}$$

with r_k a polynomial defined on I given by:

$$r_k(s) = (j-1)!\int_0^1 \frac{t^{i-1}}{(i-1)!}\frac{(1-t)^{j-i-1}}{(j-i-1)!}\varphi_{i,k}(st)\,dt.$$

Notice that r_k is a polynomial of degree at most k as $\varphi_{i,k}$ is a polynomial with degree k. Therefore if $k < \ell$, we have $\int_\triangle \varphi_{[i],k}\varphi_{[j],\ell} = 0$ since $\varphi_{j,\ell}$ is orthogonal (with respect to the measure q_j) to any polynomial of degree less than ℓ. Similar calculations show that if $k > \ell$, the integral is also 0.

Let us consider now the case $k = \ell$. We compute the coefficient ν_k of t^k in the polynomial r_k. We deduce from (6.25) that the leading coefficient $\omega_{i,k}$ of $\varphi_{i,k}$ is given by:

$$\omega_{i,k} = \rho_{i,k}\sqrt{(d-i)!(i-1)!}\,\omega'_{i,k} = \rho_{i,k}\sqrt{(d-i)!(i-1)!}\,\frac{(2k+d-1)!}{k!(k+d-1)!}.$$

Using this we obtain for ν_k :

$$\begin{aligned}\nu_k &= (j-1)!\omega_{i,k}\int_0^1 \frac{t^{k+i-1}}{(i-1)!}\frac{(1-t)^{j-i-1}}{(j-i-1)!}\,dt \\ &= \omega_{i,k}\frac{(k+i-1)!(j-1)!}{(k+j-1)!(i-1)!},\end{aligned}$$

and thus r_k has degree k. The orthonormality of $(\varphi_{j,k}, k \in \mathbb{N})$ ensures that $\int_I r_k\varphi_{j,k}\, q_j = \nu_k/\omega_{j,k}$. Therefore, we obtain:

$$\int_\triangle \varphi_{[i],k}\varphi_{[j],k} = \frac{\nu_k}{\omega_{j,k}} = \sqrt{\frac{(j-1)!(d-i)!}{(i-1)!(d-j)!}}\sqrt{\frac{(k+d-j)!(k+i-1)!}{(k+d-i)!(k+j-1)!}}.$$

Since $(j-1)!/(i-1)! \leq (k+j-1)!/(k+i-1)!$, and $(d-i)!/(d-j)! \leq (k+d-i)!/(k+d-j)!$, we can conclude that $0 \leq \int_{\triangle} \varphi_{[i],k}\varphi_{[j],k} \leq 1$.

□

This shows that the family of functions $\varphi = (\varphi_{i,k}, 1 \leq i \leq d, k \in \mathbb{N})$ is not orthogonal with respect to the Lebesgue measure on $\triangle$. For $k \in \mathbb{N}^*$, let us consider the matrix $R_k \in \mathbb{R}^{d\times d}$ with elements:

$$R_k(i,j) = \int_{\triangle} \varphi_{[i],k}\varphi_{[j],k}. \tag{6.29}$$

If $Y = (Y_1, \ldots, Y_d)$ is uniformly distributed on $\triangle$, then R_k is the correlation matrix of the random variable $(\varphi_{1,k}(Y_1), \ldots, \varphi_{d,k}(Y_d))$. Therefore it is symmetric and positive semi-definite. Let $\lambda_{k,1} \leq \ldots \leq \lambda_{k,d}$ denote the eigenvalues of R_k. We aim to find a lower bound for these eigenvalues which is independent of k.

Lemma 6.15. *For $k \in \mathbb{N}^*$, the smallest eigenvalue $\lambda_{k,d}$ of R_k is given by:*

$$\lambda_{k,d} = \frac{k}{k+d-1},$$

and we have $\lambda_{k,d} \geq 1/d$.

Proof. It is easy to check that the inverse R_k^{-1} of R_k exists and is symmetric tridiagonal with diagonal entries D_i, $1 \leq i \leq d$ and lower (and upper) diagonal elements Q_i, $1 \leq i \leq d-1$ given by:

$$D_i = \frac{(k+d-1)(k+1)+2(i-1)(d-i)}{k(k+d)} \quad \text{and} \quad Q_i = -\frac{\sqrt{i(d-i)(k+i)(k+d-i)}}{k(k+d)}.$$

The matrix R_k^{-1} is positive definite, since all of its principal minors have a positive determinant. In particular, this ensures that the eigenvalues of R_k and R_k^{-1} are all positive. Let $c_i(\lambda)$, $1 \leq i \leq d$ denote the i-th leading principal minor of the matrix $R_k^{-1} - \lambda I_d$, where I_d is the d-dimensional identity matrix. The eigenvalues of R_k^{-1} are exactly the roots of the characteristic polynomial $c_d(\lambda)$. Since R_k^{-1} is symmetric and tridiagonal, we have the following recurrence relation for $c_i(\lambda)$, $1 \leq i \leq d$:

$$c_i(\lambda) = (D_i - \lambda)c_{i-1}(\lambda) - Q_{i-1}^2 c_{i-2}(\lambda),$$

with initial values $c_0(\lambda) = 1$, $c_{-1}(\lambda) = 0$.

Let M_k be the symmetric tridiagonal matrix $d \times d$ with diagonal entries D_i, $1 \leq i \leq d$ and lower (and upper) diagonal elements $|Q_i|$, $1 \leq i \leq d-1$. Notice the characteristic polynomial of M_k is also $c_d(\lambda)$. So M_k and R_k^{-1} have the same eigenvalues.

It is easy to check that $\lambda^* = (k+d-1)/k$ is an eigenvalue of M_k with corresponding eigenvector $v = (v_1, \ldots, v_d)$ given by, for $1 \leq i \leq d$:

$$v_i = \sqrt{\frac{(d-1)!}{(d-i)!}\frac{(k+d-1)!}{(k+d-i)!}\frac{k!}{(k+i-1)!}\frac{1}{(i-1)!}}.$$

(One can check that $v' = (v'_1, \ldots, v'_d)$, with $v'_i = (-1)^{i-1}v_i$, is an eigenvector of R_k^{-1} with eigenvalue λ^*.)

The matrix M_k has non-negative elements, with positive elements in the diagonal, sub- and superdiagonal. Therefore M_k is irreducible, and we can apply the Perron-Frobenius theorem for non-negative, irreducible matrices: the largest eigenvalue of M_k has multiplicity one and is the only eigenvalue with corresponding eigenvector x such that $x > 0$. Since $v > 0$, we deduce that λ^* is the largest eigenvalue of M_k. It is also the largest eigenvalue of R_k^{-1}. Thus $1/\lambda^* = k/(k+d-1)$ is the lowest eigenvalue of R_k.

Since $\lambda_{k,d}$ is increasing in k, we have the uniform lower bound $1/d$. □

Remark 6.16. We conjecture that the eigenvalues $\lambda_{k,i}$ of R_k are given by, for $1 \leq i \leq d$:

$$\lambda_{k,i} = \frac{k(k+d)}{(k+i)(k+i-1)}.$$

6.6.4 Bounds between different norms

In this Section, we will give inequalities between different types of norms for functions defined on the simplex $\triangle$. These inequalities are used during the proof of Theorem 6.4. Let $m = (m_1, \dots, m_d) \in (\mathbb{N}^*)^d$. Recall the notation φ_m and $\theta \cdot \varphi_m$ with $\theta = (\theta_{i,k}; 1 \leq k \leq m_i, 1 \leq i \leq d) \in \mathbb{R}^{|m|}$ from Section 6.3.

For $1 \leq i \leq d$, we set $\theta_i = (\theta_{i,k}, 1 \leq k \leq m_i) \in \mathbb{R}^{m_i}$, $\varphi_{i,m} = (\varphi_{i,k}, 1 \leq k \leq m_i)$ and:

$$\theta_i \cdot \varphi_{i,m} = \sum_{k=1}^{m_i} \theta_{i,k}\varphi_{i,k} \quad \text{and} \quad \theta_i \cdot \varphi_{[i],m} = \sum_{k=1}^{m_i} \theta_{i,k}\varphi_{[i],k},$$

with $\varphi_{[i],m} = (\varphi_{[i],k}, 1 \leq k \leq m_i)$. In particular, we have $\varphi_m = \sum_{i=1}^d \varphi_{[i],m}$ and $\theta \cdot \varphi_m = \sum_{i=1}^d \theta_i \cdot \varphi_{[i],m}$. We first give lower and upper bounds on $\| \theta \cdot \varphi_m \|_{L^2}$.

Lemma 6.17. *For all $\theta \in \mathbb{R}^{|m|}$ we have:*

$$\frac{\| \theta \|}{\sqrt{d}} \leq \| \theta \cdot \varphi_m \|_{L^2} \leq \sqrt{d} \| \theta \| .$$

Proof. For the upper bound, one simply has, by the triangle inequality and the orthonormality:

$$\| \theta \cdot \varphi_m \|_{L^2} \leq \sum_{i=1}^d \| \theta_i \cdot \varphi_{i,m} \|_{L^2(q_i)} = \sum_{i=1}^d \| \theta_i \| \leq \sqrt{d} \| \theta \| .$$

For the lower bound, we have:

$$\| \theta \cdot \varphi_m \|_{L^2}^2 = \sum_{i=1}^d \sum_{k=1}^{m_i} \theta_{i,k}^2 + 2 \sum_{i<j} \sum_{k=1}^{\min(m_i,m_j)} \theta_{i,k}\theta_{j,k} \int_{\triangle} \varphi_{[i],k}\varphi_{[j],k}, \tag{6.30}$$

where we used the normality of $\varphi_{[i],k}$ with respect to the Lebesgue measure on $\triangle$ and Lemma 6.14 for the cross products. We can rewrite this in a matrix form:

$$\| \theta \cdot \varphi_m \|_{L^2}^2 \geq \sum_{k=1}^{\max(m)} (\theta_k^*)^T R_k \theta_k^*,$$

where $R_k \in \mathbb{R}^{d \times d}$ is given by (6.29) and $\theta_k^* = (\theta_{1,k}^*, \dots, \theta_{d,k}^*) \in \mathbb{R}^d$ is defined, for $1 \leq i \leq d$, $1 \leq k \leq \max(m)$, as:

$$\theta_{i,k}^* = \theta_{i,k} \mathbf{1}_{\{k \leq m_i\}}.$$

Since, according to Lemma 6.15, all the eigenvalues of R_k are uniformly larger than $1/d$, this gives:

$$\| \theta \cdot \varphi_m \|_{L^2}^2 \geq \frac{1}{d} \sum_{k=1}^{\max(m)} \| \theta_k^* \|^2 = \frac{\| \theta \|^2}{d}.$$

This concludes the proof. □

We give an inequality between different norms for polynomials defined on I.

Lemma 6.18. *If h is a polynomial of degree less then or equal to n on I, then we have for all $1 \leq i \leq d$:*

$$\| h \|_\infty \leq \sqrt{2(d-1)!}(n+d)^d \| h \|_{L^2(q_i)}$$

Proof. There exists $(\beta_k, 0 \leq k \leq n)$ such that $h = \sum_{k=0}^n \beta_k \varphi_{i,k}$. By the Cauchy-Schwarz inequality, we have:

$$|h| \leq \left(\sum_{k=0}^n \beta_k^2 \right)^{1/2} \left(\sum_{k=0}^n \varphi_{i,k}^2 \right)^{1/2}. \tag{6.31}$$

We deduce from Definition 6.12 of $\varphi_{i,k}$ and (6.27) that:

$$\| \varphi_{i,k} \|_\infty = \sqrt{\frac{(2k+d)(k+d-1)!}{k!}} \max\left(\sqrt{\frac{(i-1)!(k+d-i)!}{(d-i)!(k+i-1)!}}, \sqrt{\frac{(d-i)!(k+i-1)!}{(i-1)!(k+d-i)!}} \right).$$

For all $1 \leq i \leq d$, we have the uniform upper bound:

$$\| \varphi_{i,k} \|_\infty \leq \sqrt{(d-1)!}\sqrt{2k+d}\frac{(k+d-1)!}{k!}. \tag{6.32}$$

This implies that for $t \in I$:

$$\sum_{k=0}^{n} \varphi_{i,k}^2(t) \leq \sum_{k=0}^{n} \| \varphi_{i,k}^2 \|_\infty \leq (d-1)! \sum_{k=0}^{n} (2k+d) \left(\frac{(k+d-1)!}{k!} \right)^2 \leq 2(d-1)!(n+d)^{2d}.$$

Bessel's inequality implies that $\sum_{k=0}^{n} \beta_k^2 \leq \| h \|_{L^2(q_i)}^2$. We conclude the proof using (6.31). □

We recall the notation S_m of the linear space spanned by $(\varphi_{[i],k}; 1 \leq k \leq m_i, 1 \leq i \leq d)$, and the different norms introduced in Section 6.7.

Lemma 6.19. *Let $m \in (\mathbb{N}^*)^d$ and $\kappa_m = \sqrt{2d!}\sqrt{\sum_{i=1}^{d}(m_i+d)^{2d}}$. Then we have for every $g \in S_m$: $\| g \|_\infty \leq \kappa_m \| g \|_{L^2}$.*

Proof. Let $g \in S_m$. We can write $g = \theta \cdot \varphi_m$ for a unique $\theta \in \mathbb{R}^{|m|}$. Let $g_i = \theta_i \cdot \varphi_{i,m}$ so that $g = \sum_{i=1}^{d} g_{[i]}$, where g_i is a polynomial defined on I of degree at most m_i for all $1 \leq i \leq d$. We have:

$$\begin{aligned}
\| g \|_\infty &\leq \sum_{i=1}^{d} \| g_i \|_\infty \\
&\leq \sqrt{2(d-1)!} \sum_{i=1}^{d} (m_i+d)^d \| g_i \|_{L^2(q_i)} \\
&\leq \frac{\kappa_m}{\sqrt{d}} \left(\sum_{i=1}^{d} \| g_i \|_{L^2(q_i)}^2 \right)^{1/2} \\
&= \frac{\kappa_m}{\sqrt{d}} \| \theta \| \\
&\leq \kappa_m \| \theta \cdot \varphi_m \|_{L^2} \\
&= \kappa_m \| g \|_{L^2}.
\end{aligned}$$

where we used Lemma 6.18 for the second inequality, Cauchy-Schwarz for the third inequality, and Lemma 6.17 for the fourth inequality. □

Remark 6.20. For d fixed, κ_m as a function of m verifies:

$$\kappa_m = O\left(\sqrt{\sum_{i=1}^{d} m_i^{2d}} \right) = O(|m|^d).$$

6.6.5 Bounds on approximations

Now we bound the L^2 and L^∞ norm of the approximation error of additive functions where each component belongs to a Sobolev space. Let $m = (m_1, \ldots, m_d) \in (\mathbb{N}^*)^d$, $r = (r_1, \ldots, r_d) \in (\mathbb{N}^*)^d$ such that $m_i + 1 \geq r_i$ for all $1 \leq i \leq d$. Let $\ell = \sum_{i=1}^{d} \ell_{[i]}$ with $\ell_i \in W_{r_i}^2(q_i)$ and $\int_I \ell_i q_i = 0$ for $1 \leq i \leq d$. Let ℓ_{i,m_i} be the orthogonal projection in $L^2(q_i)$ of ℓ_i on the span of $(\varphi_{i,k}, 0 \leq k \leq m_i)$ given by $\ell_{i,m_i} = \sum_{k=1}^{m_i} \left(\int_I \ell_i \varphi_{i,k} q_i \right) \varphi_{i,k}$. Then $\ell_m = \sum_{i=1}^{d} \ell_{[i],m_i}$ is the approximation of ℓ on S_m given by (6.47). We start by giving a bound on the $L^2(q_i)$ norm of the error when we approximate ℓ_i by ℓ_{i,m_i}.

Lemma 6.21. *For each $1 \le i \le d$, $m_i + 1 \ge r_i$ and $\ell_i \in W_{r_i}^2(q_i)$, we have:*

$$\|\ell_i - \ell_{i,m_i}\|_{L^2(q_i)}^2 \le \frac{2^{-2r_i}(m_i+1-r_i)!(m_i+d)!}{(m_i+1)!(m_i+d+r_i)!}\,\|\ell_i^{(r_i)}\|_{L^2(q_i)}^2 . \tag{6.33}$$

Proof. Notice that (6.26) implies that the series $(\varphi_{i,k}^{(r_i)}, k \ge r_i)$ is orthogonal on I with respect to the weight function $v_i(t) = (1-t)^{d-i+r_i} t^{i-1+r_i}$, and the normalizing constants $\kappa_{i,k} \ge 0$ are given by:

$$\begin{aligned}
\kappa_{i,k}^2 &= \int_0^1 \left(\varphi_{i,k}^{(r_i)}(t)\right)^2 v_i(t)\,dt \\
&= \rho_{i,k}^2 (d-i)!(i-1)! \int_0^1 \left(\frac{d^{r_i}}{dt^{r_i}} P_k^{(d-i,i-1)}(2t-1)\right)^2 v_i(t)\,dt \\
&= \rho_{i,k}^2 (d-i)!(i-1)! \frac{((k+d-1+r_i)!)^2}{2^{d+2r_i}((k+d-1)!)^2} \int_{-1}^1 \left(P_{k-r_i}^{(d-i+r_i,i-1+r_i)}(s)\right)^2 w_{d-i+r_i,i-1+r_i}(s)\,ds \\
&= (d-i)!(i-1)! \frac{k!(k+d-1+r_i)!}{(k-r_i)!(k+d-1)!},
\end{aligned} \tag{6.34}$$

where we used the definition of $\varphi_{i,k}$ for the second equality, (6.26) for the third equality and (6.24) for the fourth equality. Notice that $\kappa_{i,k}$ is non-decreasing as a function of k. Since $\ell_i - \ell_{i,m_i} = \sum_{k=m_i+1}^{\infty} \beta_{i,k}\varphi_{i,k}$, we have:

$$\|\ell_i - \ell_{i,m_i}\|_{L^2(q_i)}^2 = \sum_{k=m_i+1}^{\infty} \beta_{i,k}^2 \le \frac{1}{\kappa_{i,m_i+1}^2} \sum_{k=m_i+1}^{\infty} \kappa_{i,k}^2 \beta_{i,k}^2 \le \frac{1}{\kappa_{i,m_i+1}^2} \sum_{k=r_i}^{\infty} \kappa_{i,k}^2 \beta_{i,k}^2, \tag{6.35}$$

where the first inequality is due to the monotonicity of $\kappa_{i,k}$ as k increases. Thanks to (6.26) and the definition of $\kappa_{i,k}$, we get that $(\varphi_{i,k}^{(r_i)}/\kappa_{i,k}, k \ge r_i)$ is an orthonormal basis of $L^2(v_i)$. Therefore, we have

$$\sum_{k=r_i}^{\infty} \kappa_{i,k}^2 \beta_{i,k}^2 = \int_0^1 \left(\ell_i^{(r_i)}(t)\right)^2 v_i(t)\,dt \le \frac{(d-i)!(i-1)!}{2^{2r_i}} \|\ell_i^{(r_i)}\|_{L^2(q_i)}^2, \tag{6.36}$$

since $\sup_{t\in I} q_i(t)/v_i(t) = (d-i)!(i-1)!/2^{2r_i}$. This and (6.35) implies (6.33). □

Lemma 6.21 yields a simple bound on the L^2 norm of the approximation error $\ell - \ell_m$.

Corollary 6.22. *For $m = (m_1, \dots, m_d)$, $m_i + 1 \ge r_i$ and $\ell_i \in W_{r_i}^2(q_i)$ for all $1 \le i \le d$, we get:*

$$\|\ell - \ell_m\|_{L^2} = O\left(\sqrt{\sum_{i=1}^d m_i^{-2r_i}}\right).$$

Proof. We have:

$$\|\ell - \ell_m\|_{L^2} \le \sum_{i=1}^d \|\ell_i - \ell_{i,m_i}\|_{L^2(q_i)} = O\left(\sum_{i=1}^d m_i^{-r_i}\right) = O\left(\sqrt{\sum_{i=1}^d m_i^{-2r_i}}\right),$$

where we used (6.33) for the first equality. □

Lastly, we bound the L^∞ norm of the approximation error.

Lemma 6.23. *For each $1 \le i \le d$, $m_i + 1 \ge r_i > d$ and $\ell_i \in W_{r_i}^2(q_i)$, we have:*

$$\|\ell_i - \ell_{i,m_i}\|_\infty \le \frac{2^{-r_i}\sqrt{2(d-1)!}\,e^{r_i}}{\sqrt{2r_i - 2d - 1}} \frac{1}{(m_i + r_i)^{r_i - d - \frac{1}{2}}} \|\ell_i^{(r_i)}\|_{L^2(q_i)} . \tag{6.37}$$

Proof. We recall the constants $\kappa_{i,k}$, $1 \leq i \leq d$, $1 \leq k \leq m_i$ given by (6.34). Since $\ell_i - \ell_{i,m_i} = \sum_{k=m_i+1}^{\infty} \beta_{i,k}\varphi_{i,k}$ we have:

$$\begin{aligned}
\| \ell_i - \ell_{i,m_i} \|_\infty &= \left\| \sum_{k=m_i+1}^{\infty} \beta_{i,k}\varphi_{i,k} \right\|_\infty \\
&\leq \sum_{k=m_i+1}^{\infty} |\beta_{i,k}| \, \|\varphi_{i,k}\|_\infty \\
&\leq \sqrt{\sum_{k=m_i+1}^{\infty} \frac{\|\varphi_{i,k}\|_\infty^2}{\kappa_{i,k}^2}} \sqrt{\sum_{k=m_i+1}^{\infty} \kappa_{i,k}^2 \beta_{i,k}^2} \\
&\leq \sqrt{\sum_{k=m_i+1}^{\infty} \frac{2(d-1)!(k+d)^{2d}}{\kappa_{i,k}^2}} \sqrt{\frac{(d-i)!(i-1)!}{2^{2r_i}}} \, \| \ell_i^{(r_i)} \|_{L^2(q_i)} \\
&\leq \sqrt{\sum_{k=m_i+1}^{\infty} \frac{2(d-1)!}{(d-i)!(i-1)!} \frac{\mathrm{e}^{2r_i}}{(k+r_i)^{2r_i-2d}}} \sqrt{\frac{(d-i)!(i-1)!}{2^{2r_i}}} \, \| \ell_i^{(r_i)} \|_{L^2(q_i)} \\
&\leq \frac{2^{-r_i}\sqrt{2(d-1)!}\,\mathrm{e}^{r_i}}{\sqrt{2r_i-2d-1}\sqrt{(m_i+r_i)^{2r_i-2d-1}}} \, \| \ell_i^{(r_i)} \|_{L^2(q_i)},
\end{aligned}$$

where we used Cauchy-Schwarz for the second inequality, (6.32) and (6.36) for the third inequality, $\kappa_{i,k}^2 \geq (d-i)!(i-1)!(k+r_i)^{2r_i}\,\mathrm{e}^{-2r_i}$ for the fourth inequality, and $\sum_{k=m_i+1}^{\infty}(k+r_i)^{-2r_i+2d} \leq (2r_i-2d-1)^{-1}(m_i+r_i)^{-2r_i+2d+1}$ for the fifth inequality. □

Corollary 6.24. *There exists a constant $\mathcal{C} > 0$ such that for all $\ell_i \in W_{r_i}^2(q_i)$ and $m_i+1 \geq r_i > d$ for all $1 \leq i \leq d$, we have:*

$$\| \ell - \ell_m \|_\infty \leq \mathcal{C} \sum_{i=1}^{d} \| \ell_i^{(r_i)} \|_{L^2(q_i)} .$$

Proof. Notice that for $m_i + 1 \geq r_i > d$, we have:

$$\frac{2^{-r_i}\sqrt{2(d-1)!}\,\mathrm{e}^{r_i}}{\sqrt{2r_i-2d-1}} \frac{1}{(m_i+r_i)^{r_i-d-\frac{1}{2}}} \leq \frac{2^{-r_i}\sqrt{2(d-1)!}\,\mathrm{e}^{r_i}}{\sqrt{2r_i-2d-1}} \frac{1}{(2r_i-1)^{r_i-d-\frac{1}{2}}},$$

and that the right hand side is bounded by a constant $\mathcal{C} > 0$ for all $r_i \in \mathbb{N}^*$. Therefore:

$$\| \ell - \ell_m \|_\infty \leq \sum_{i=1}^{d} \| \ell_i - \ell_{i,m_i} \|_\infty \leq \mathcal{C} \sum_{i=1}^{d} \| \ell_i^{(r_i)} \|_{L^2(q_i)} .$$

□

6.7 Preliminary elements for the proof of Theorem 6.4

We adapt the results from [12] to our setting. Let us recall Lemmas 1 and 2 of [12].

Lemma 6.25 (Lemma 1 of [12])**.** *Let $g, h \in \mathcal{P}(\triangle)$. If $\| \log(g/h) \|_\infty < +\infty$, then we have:*

$$D(g\|h) \geq \frac{1}{2} \mathrm{e}^{-\|\log(g/h)\|_\infty} \int_\triangle g \log^2(g/h) , \tag{6.38}$$

and for any $\kappa \in \mathbb{R}$:

$$D(g\|h) \leq \frac{1}{2} \mathrm{e}^{\|\log(g/h)-\kappa\|_\infty} \int_\triangle g \left(\log(g/h) - \kappa\right)^2 , \tag{6.39}$$

$$\int_\triangle \frac{(g-h)^2}{g} \leq \mathrm{e}^{2\left(\|\log(g/h)-\kappa\|_\infty - \kappa\right)} \int_\triangle g \left(\log(g/h) - \kappa\right)^2 . \tag{6.40}$$

Lemma 6.25 readily implies the following Corollary.

Corollary 6.26. *Let $g, h \in \mathcal{P}(\triangle)$. If $\|\log(g/h)\|_\infty < +\infty$, then we have, for any constant $\kappa \in \mathbb{R}$:*

$$D(g\|h) \leq \frac{1}{2} e^{\|\log(g/h)-\kappa\|_\infty} \|g\|_\infty \int_\triangle (\log(g/h) - \kappa)^2, \tag{6.41}$$

and:

$$\|g - h\|_{L^2} \leq \|g\|_\infty e^{(\|\log(g/h)-\kappa\|_\infty - \kappa)} \|\log(g/h) - \kappa\|_{L^2}. \tag{6.42}$$

Recall Definition 6.1 for densities f^0 with a product form on $\triangle$. We give a few bounds between the L^∞ norms of $\log(f^0)$, ℓ^0 and the constant a_0.

Lemma 6.27. *Let $f^0 \in \mathcal{P}(\triangle)$ given by Definition 6.1. Then we have:*

$$|a_0| \leq \|\ell^0\|_\infty + |\log(d!)|, \qquad \|\log(f^0)\|_\infty \leq 2\|\ell^0\|_\infty + |\log(d!)|, \tag{6.43}$$

$$|a_0| \leq \|\log(f^0)\|_\infty, \qquad \|\ell^0\|_\infty \leq 2\|\log(f^0)\|_\infty. \tag{6.44}$$

Proof. The first part of (6.43) can be obtained by bounding ℓ^0 with $\|\ell^0\|_\infty$ in the definition of a_0. The second part is a direct consequence of this. The first part of (6.44) can be deduced from the fact that $\int_\triangle \ell^0 = 0$. The second part is again a direct consequence of the first part. □

Let $m \in (\mathbb{N}^*)^d$. Recall the application A_m defined in (6.7) and set $\Omega_m = A_m(\mathbb{R}^{|m|})$. For $\alpha \in \mathbb{R}^{|m|}$, we define the function $\mathscr{F}_\alpha$ on $\mathbb{R}^{|m|}$ by:

$$\mathscr{F}_\alpha(\theta) = \theta \cdot \alpha - \psi(\theta). \tag{6.45}$$

Recall also the log-additive exponential model f_θ given by (6.6).

Lemma 6.28 (Lemma 3 of [12])**.** *Let $m \in (\mathbb{N}^*)^d$. The application A_m is one-to-one from $\mathbb{R}^{|m|}$ onto Ω_m, with inverse say Θ_m. Let $f \in \mathcal{P}(\triangle)$ such that $\alpha = \int_\triangle \varphi_m f$ belongs to Ω_m. Then for all $\theta \in \mathbb{R}^{|m|}$, we have with $\theta^* = \Theta_m(\alpha)$:*

$$D(f\|f_\theta) = D(f\|f_{\theta^*}) + D(f_{\theta^*}\|f_\theta). \tag{6.46}$$

Furthermore, θ^* achieves $\max_{\theta\in\mathbb{R}^{|m|}} \mathscr{F}_\alpha(\theta)$ as well as $\min_{\theta\in\mathbb{R}^{|m|}} D(f\|f_\theta)$.

Definition 6.29. Let $m \in (\mathbb{N}^*)^d$. For $f \in \mathcal{P}(\triangle)$ such that $\alpha = \int_\triangle \varphi_m f \in \Omega_m$, the probability density f_{θ^*}, with $\theta^* = \Theta_m(\alpha)$ (that is $\int_\triangle \varphi_m f = \int_\triangle \varphi_m f_{\theta^*}$), is called the information projection of f.

The information projection of a density f is the closest density in the exponential family (6.6) with respect to the Kullback-Leibler distance to f.

We consider the linear space of real valued functions defined on $\triangle$ and generated by φ_m:

$$S_m = \{\theta \cdot \varphi_m; \theta \in \mathbb{R}^{|m|}\}. \tag{6.47}$$

Let $\kappa_m = \sqrt{2d!}\sqrt{\sum_{i=1}^d (m_i + d)^{2d}}$. The following Lemma summarizes Lemmas 6.17 and 6.19.

Lemma 6.30. *Let $m \in (\mathbb{N}^*)^d$. We have for all $g \in S_m$:*

$$\|g\|_\infty \leq \kappa_m \|g\|_{L^2}, \tag{6.48}$$

For all $\theta \in \mathbb{R}^{|m|}$, we have:

$$\frac{\|\theta\|}{\sqrt{d}} \leq \|\theta \cdot \varphi_m\|_{L^2} \leq \sqrt{d}\|\theta\|. \tag{6.49}$$

Now we give upper and lower bounds for the Kullback-Leibler distance between two members of the exponential family f_θ and $f_{\theta'}$ in terms of the Euclidean distance $\|\theta - \theta'\|$. Notice that for all $\theta \in \mathbb{R}^{|m|}$, $\|\log(f_\theta)\|_\infty = \sup_{x\in\triangle} |\log(f_\theta(x))|$ is finite.

Lemma 6.31. *Let $m \in (\mathbb{N}^*)^d$. For $\theta, \theta' \in \mathbb{R}^{|m|}$, we have:*

$$\| \log(f_\theta / f_{\theta'}) \|_\infty \leq 2\sqrt{d}\, \kappa_m \, \| \theta - \theta' \|, \tag{6.50}$$

$$D(f_\theta \| f_{\theta'}) \leq \frac{d}{2}\, \mathrm{e}^{\| \log(f_\theta) \|_\infty + \sqrt{d}\, \kappa_m \, \| \theta - \theta' \|} \, \| \theta - \theta' \|^2, \tag{6.51}$$

$$D(f_\theta \| f_{\theta'}) \geq \frac{1}{2d}\, \mathrm{e}^{- \| \log(f_\theta) \|_\infty - 2\sqrt{d}\, \kappa_m \, \| \theta - \theta' \|} \, \| \theta - \theta' \|^2 . \tag{6.52}$$

Proof. Since $\psi(\theta') - \psi(\theta) = \log\left(\int_\triangle \mathrm{e}^{(\theta' - \theta)\cdot \varphi_m} f_\theta \right)$, we get $|\psi(\theta') - \psi(\theta)| \leq \| (\theta' - \theta) \cdot \varphi_m \|_\infty$. This implies that:

$$\begin{aligned} \| \log(f_\theta / f_{\theta'}) \|_\infty &\leq 2 \, \| (\theta - \theta') \cdot \varphi_m \|_\infty \\ &\leq 2\kappa_m \, \| (\theta - \theta') \cdot \varphi_m \|_{L^2} \\ &\leq 2\sqrt{d}\, \kappa_m \, \| \theta - \theta' \|, \end{aligned}$$

where we used (6.6) for the first inequality, (6.48) for the second and (6.49) for the third. To prove (6.51), we use (6.39) with $\kappa = \psi(\theta') - \psi(\theta)$. This gives:

$$\begin{aligned} D(f_\theta \| f_{\theta'}) &\leq \frac{1}{2}\, \mathrm{e}^{\| (\theta - \theta')\cdot \varphi_m \|_\infty} \int_\triangle f_\theta \left((\theta - \theta') \cdot \varphi_m \right)^2 \\ &\leq \frac{1}{2}\, \mathrm{e}^{\| \log(f_\theta) \|_\infty + \sqrt{d}\, \kappa_m \, \| \theta - \theta' \|} \, \| (\theta - \theta') \cdot \varphi_m \|_{L^2}^2 \\ &\leq \frac{d}{2}\, \mathrm{e}^{\| \log(f_\theta) \|_\infty + \sqrt{d}\, \kappa_m \, \| \theta - \theta' \|} \, \| \theta - \theta' \|^2, \end{aligned}$$

where we used (6.48) and (6.49) for the second inequality, and (6.49) for the third. To prove (6.52), we use (6.38). We obtain:

$$\begin{aligned} D(f_\theta \| f_{\theta'}) &\geq \frac{1}{2}\, \mathrm{e}^{- \| \log(f_\theta / f_{\theta'}) \|_\infty} \int_\triangle f_\theta \left((\theta - \theta') \cdot \varphi_m - (\psi(\theta) - \psi(\theta')) \right)^2 \\ &\geq \frac{1}{2}\, \mathrm{e}^{- \| \log(f_\theta) \|_\infty - 2\sqrt{d}\, \kappa_m \, \| \theta - \theta' \|} \int_\triangle \left((\theta - \theta') \cdot \varphi_m - (\psi(\theta) - \psi(\theta')) \right)^2 \\ &\geq \frac{1}{2}\, \mathrm{e}^{- \| \log(f_\theta) \|_\infty - 2\sqrt{d}\, \kappa_m \, \| \theta - \theta' \|} \, \| (\theta - \theta') \cdot \varphi_m \|_{L^2}^2 \\ &\geq \frac{1}{2d}\, \mathrm{e}^{- \| \log(f_\theta) \|_\infty - 2\sqrt{d}\, \kappa_m \, \| \theta - \theta' \|} \, \| \theta - \theta' \|^2, \end{aligned}$$

where we used (6.50) for the second inequality, the fact that the functions $(\varphi_{[i],k}, 1 \leq i \leq d, 1 \leq k \leq m_i)$ are orthogonal to the constant function with respect to the Lebesgue measure on $\triangle$ for the third inequality, and (6.49) for the fourth inequality. □

Now we will show that the application Θ_m is locally Lipschitz.

Lemma 6.32. *Let $m \in (\mathbb{N}^*)^d$ and $\theta \in \mathbb{R}^{|m|}$. If $\alpha \in \mathbb{R}^{|m|}$ satisfies:*

$$\| A_m(\theta) - \alpha \| \leq \frac{\mathrm{e}^{-(1 + \| \log(f_\theta) \|_\infty)}}{6 d^{\frac{3}{2}} \kappa_m}, \tag{6.53}$$

Then α belongs to Ω_m and $\theta^* = \Theta_m(\alpha)$ exists. Let τ be such that:

$$6 d^{\frac{3}{2}}\, \mathrm{e}^{1 + \| \log(f_\theta) \|_\infty}\, \kappa_m \, \| A_m(\theta) - \alpha \| \leq \tau \leq 1.$$

Then θ^* satisfies:

$$\| \theta - \theta^* \| \leq 3d\, \mathrm{e}^{\tau + \| \log(f_\theta) \|_\infty} \, \| A_m(\theta) - \alpha \|, \tag{6.54}$$

$$\| \log(f_\theta / f_{\theta^*}) \|_\infty \leq 6 d^{\frac{3}{2}}\, \mathrm{e}^{\tau + \| \log(f_\theta) \|_\infty}\, \kappa_m \, \| A_m(\theta) - \alpha \| \leq \tau, \tag{6.55}$$

$$D(f_\theta \| f_{\theta^*}) \leq 3d\, \mathrm{e}^{\tau + \| \log(f_\theta) \|_\infty} \, \| A_m(\theta) - \alpha \|^2 . \tag{6.56}$$

Proof. Suppose that $\alpha \neq A_m(\theta)$ (otherwise the results are trivial). Recall $\mathscr{F}_\alpha$ defined in (6.45). We have, for all $\theta' \in \mathbb{R}^{|m|}$:

$$\begin{aligned}\mathscr{F}_\alpha(\theta) - \mathscr{F}_\alpha(\theta') &= (\theta - \theta') \cdot \alpha + \psi(\theta') - \psi(\theta) \\ &= D\left(f_\theta \| f_{\theta'}\right) - (\theta - \theta') \cdot (A_m(\theta) - \alpha). \end{aligned} \tag{6.57}$$

Using (6.52) and the Cauchy-Schwarz inequality, we obtain the strict inequality:

$$\mathscr{F}_\alpha(\theta) - \mathscr{F}_\alpha(\theta') > \frac{1}{3d} \mathrm{e}^{-\|\log(f_\theta)\|_\infty - 2\sqrt{d}\,\kappa_m \|\theta - \theta'\|} \|\theta - \theta'\|^2 - \|\theta - \theta'\| \, \|A_m(\theta) - \alpha\| .$$

We consider the ball centered at θ: $B_r = \{\theta' \in \mathbb{R}^{|m|}, \|\theta - \theta'\| \leq r\}$ with radius r given by $r = 3d\,\mathrm{e}^{\tau + \|\log(f_\theta)\|_\infty} \|A_m(\theta) - \alpha\|$. For all $\theta' \in \partial B_r$, we have:

$$\mathscr{F}_\alpha(\theta) - \mathscr{F}_\alpha(\theta') > \left(\mathrm{e}^{\tau - 6d^{\frac{3}{2}} \kappa_m \|A_m(\theta) - \alpha\| \, \mathrm{e}^{\tau + \|\log(f_\theta)\|_\infty}} - 1\right) 3d\,\mathrm{e}^{\tau + \|\log(f_\theta)\|_\infty} \|A_m(\theta) - \alpha\|^2 .$$

The right hand side is non-negative as $6d^{\frac{3}{2}}\,\mathrm{e}^{1 + \|\log(f_\theta)\|_\infty}\,\kappa_m \|A_m(\theta) - \alpha\| \leq \tau \leq 1$, see the condition on τ. Thus, the value of $\mathscr{F}_\alpha$ at θ, an interior point of B_r, is larger than the values of $\mathscr{F}_\alpha$ on ∂B_r. Therefore $\mathscr{F}_\alpha$ is maximal at a point, say θ^*, in the interior of B_r. Since the gradient of $\mathscr{F}_\alpha$ at θ^* equals 0, we have $\nabla \mathscr{F}_\alpha(\theta^*) = \alpha - \int_\triangle \varphi_m f_{\theta^*} = 0$, which means that $\alpha \in \Omega_m$ and $\theta^* = \Theta_m(\alpha)$. Since θ^* is inside B_r, we get (6.54). The upper bound (6.55) is due to (6.50) of Lemma 6.31. To prove (6.56), we use (6.57) and the fact that $\mathscr{F}_\alpha(\theta) - \mathscr{F}_\alpha(\theta^*) \leq 0$, which gives:

$$D\left(f_\theta \| f_{\theta^*}\right) \leq (\theta - \theta^*) \cdot (A_m(\theta) - \alpha) \leq \|\theta - \theta^*\| \, \|A_m(\theta) - \alpha\| \leq 3d\,\mathrm{e}^{\tau + \|\log(f_\theta)\|_\infty} \|A_m(\theta) - \alpha\|^2 .$$

□

6.8 Proof of Theorem 6.4

In this Section, we first show that the information projection f_{θ^*} of f^0 onto $\{f_\theta, \theta \in \mathbb{R}^{|m|}\}$ exists for all $m \in (\mathbb{N}^*)^d$. Moreover, the maximum likelihood estimator $\hat{\theta}_{m,n}$, defined in (6.9) based on an i.i.d sample $\mathbb{X}^n$, verifies almost surely $\hat{\theta}_{m,n} = \Theta_m(\hat{\mu}_{m,n})$ for $n \geq 2$ with $\hat{\mu}_{m,n}$ the empirical mean given by (6.8). Recall $\Omega_m = A_m(\mathbb{R}^{|m|})$ with A_m defined by (6.7).

Lemma 6.33. *The mean $\alpha = \int_\triangle \varphi_m f^0$ verifies $\alpha \in \Omega_m$ and the empirical mean $\hat{\mu}_{m,n}$ verifies $\hat{\mu}_{m,n} \in \Omega_m$ almost surely when $n \geq 2$.*

Remark 6.34. By Lemma 6.28, this also means that $\hat{\theta}_{m,n} = \operatorname{argmax}_{\theta \in \mathbb{R}^{|m|}} \mathcal{F}_{\hat{\mu}_{m,n}}(\theta)$, and since $\mathcal{F}_{\hat{\mu}_{m,n}}(\theta) = (1/n) \sum_{j=1}^n \log(f_\theta(X^j))$, the estimator $\hat{f}_{m,n} = f_{\hat{\theta}_{m,n}}$ is the maximum likelihood estimator of f^0 in the model $\{f_\theta, \theta \in \mathbb{R}^m\}$ based on $\mathbb{X}^n$.

Proof. Notice that $\psi(\theta) = \log(\mathbb{E}[\exp(\theta \cdot \psi_m(U))]) - \log(d!)$, where U is a random vector uniformly distributed on $\triangle$. The Hessian matrix $\nabla^2 \psi(\theta)$ is equal to the covariance matrix of $\varphi_m(X)$, where X has density f_θ. Therefore $\nabla^2 \psi(\theta)$ is positive semi-definite, and we show that it is positive definite too. Indeed, for $\lambda \in \mathbb{R}^{|m|}$, $\lambda^T \nabla^2 \psi(\theta) \lambda = 0$ is equivalent to $\mathbb{E}[(\lambda \cdot \varphi_m(X))^2] = 0$, which implies that $\lambda \cdot \varphi_m(X) = 0$ a.e. on $\triangle$. Since $(\varphi_{i,k}, 1 \leq i \leq d, 1 \leq k \leq m_i)$ are linearly independent, this means $\lambda = 0$. Thus $\nabla^2 \psi(\theta)$ is positive definite, providing that $\theta \mapsto \psi(\theta)$ is a strictly convex function.

Let $\psi^* : \mathbb{R}^{|m|} \to \mathbb{R} \cup \{+\infty\}$ denote the Legendre-Fenchel transformation of the function $\theta \mapsto \psi(\theta)$, i.e. for $\alpha \in \mathbb{R}^{|m|}$:

$$\psi^*(\alpha) = \sup_{\theta \in \mathbb{R}^{|m|}} \alpha \cdot \theta - \psi(\theta) = \sup_{\theta \in \mathbb{R}^{|m|}} \mathcal{F}_\alpha(\theta).$$

Suppose that $\alpha \in \Omega_m$. Then according to Lemma 6.28, $\psi^*(\alpha) = \mathcal{F}_\alpha(\theta^*)$ with $\theta^* = \Theta_m(\alpha)$, thus $\psi^*(\alpha)$ is finite. Therefore $\Omega_m \subseteq \mathrm{Dom}\,(\psi^*)$, where $\mathrm{Dom}\,(\psi^*) = \{\alpha \in \mathbb{R}^{|m|} : \psi^*(\alpha) < +\infty\}$.

Inversely, let $\alpha \in \text{Dom}\,(\psi^*)$. This ensures that $\theta^* = \text{argmax}_{\theta\in\mathbb{R}^{|m|}} \mathcal{F}_\alpha(\theta)$ exists uniquely, since $\mathcal{F}_\alpha(\theta)$ is finite for all $\theta \in \mathbb{R}^{|m|}$, $\alpha \in \mathbb{R}^{|m|}$. This also implies that:

$$0 = \nabla \mathcal{F}_\alpha(\theta^*) = \alpha - \int_\triangle \varphi_m f_{\theta^*} = \alpha - A_m(\theta^*),$$

giving $\alpha \in \Omega_m$. Thus we obtain $\Omega_m = \text{Dom}\,(\psi^*)$. By Lemma 6.32, we have that Ω_m is an open subset of $\mathbb{R}^{|m|}$. Set $\Upsilon = \text{int}\,(\text{cv}\,(\text{supp}\,(\varphi_m(U))))$, where int (A) and cv (A) is the interior and convex hull of a set $A \subseteq \mathbb{R}^{|m|}$, respectively. Thanks to Lemma 4.1. of [1], we have Dom $(\psi^*) = \Upsilon$. The proof is complete as soon as we prove that $\alpha \in \Upsilon$ and $\hat{\mu}_{m,n} \in \Upsilon$ almost surely when $n \geq 2$. Since $(\varphi_{i,k}, 1 \leq i \leq d, 1 \leq k \leq m_i)$ are linearly independent polynomials, they coincide only on a finite number of points. This directly implies that $\alpha \in \Upsilon$. To show that $\hat{\mu}_{m,n} \in \Upsilon$, notice that the probability measures of $\varphi_m(X)$ and $\varphi_m(U)$ are equivalent. Therefore it is sufficient to prove that $(1/n)\sum_{j=1}^n \varphi_m(U^j) \in \Upsilon$, with $(U^1, \ldots, U^n)$ i.i.d. random vectors uniformly distributed on $\triangle$. The linear independence of $(\varphi_{i,k}, 1 \leq i \leq d, 1 \leq k \leq m_i)$ and the fact that U^j, $1 \leq j \leq n$ are uniformly distributed on $\triangle$ easily implies that for $n \geq 2$, $1/n \sum_{j=1}^n \varphi_m(U^j) \in \Upsilon$, and the proof is complete.

□

We divide the proof of Theorem 6.4 into two parts: first we bound the error due to the bias of the proposed exponential model, then we bound the error due to the variance of the sample estimation. We formulate the results in two general Propositions, which can be later specified to get Theorem 6.4.

6.8.1 Bias of the estimator

The bias error comes from the information projection of the true underlying density f^0 onto the family of the exponential series model $\{f_\theta, \theta \in \mathbb{R}^{|m|}\}$. We recall the linear space S_m spanned by $(\varphi_{[i],k}, 1 \leq k \leq m_i, 1 \leq i \leq d)$ where $\varphi_{i,k}$ is a polynomial of degree k, and the form of the probability density f^0 given in (6.5). For $1 \leq i \leq d$, let $\ell^0_{i,m}$ be the orthogonal projection in $L^2(q_i)$ of ℓ^0_i on the vector space spanned by $(\varphi_{i,k}, 0 \leq k \leq m_i)$ or equivalently on the vector space spanned by $(\varphi_{i,k}, 1 \leq k \leq m_i)$, as we assumed that $\int_I \ell^0_i q_i = 0$. We set $\ell^0_m = \sum_{i=1}^d \ell^0_{[i],m}$ the approximation of ℓ^0 on S_m. In particular we have $\ell^0_m = \theta^0 \cdot \varphi_m$ for some $\theta^0 \in \mathbb{R}^{|m|}$. Let:

$$\Delta_m = \| \ell^0 - \ell^0_m \|_{L^2} \quad \text{and} \quad \gamma_m = \| \ell^0 - \ell^0_m \|_\infty$$

denote the L^2 and L^∞ errors of the approximation of ℓ^0 by ℓ^0_m on the simplex $\triangle$.

Proposition 6.35. *Let $f^0 \in \mathcal{P}(\triangle)$ have a product form given by Definition 6.1. Let $m \in (\mathbb{N}^*)^d$. The information projection f_{θ^*} of f^0 exists (with $\theta^* \in \mathbb{R}^{|m|}$ and $\int_\triangle \varphi_m f_{\theta^*} = \int_\triangle \varphi_m f^0$) and verifies, with $\mathfrak{A}_1 = \frac{1}{2}\,\mathrm{e}^{\gamma_m + \|\log(f^0)\|_\infty}$:*

$$D\left(f^0 \| f_{\theta^*}\right) \leq \mathfrak{A}_1 \Delta_m^2. \tag{6.58}$$

Proof. The existence of θ^* is due to Lemma 6.33. Thanks to Lemma 6.28 and (6.41) with $\kappa = \psi(\theta^0) - \mathrm{a}_0$, we can deduce that:

$$D\left(f^0 \| f_{\theta^*}\right) \leq D\left(f^0 \| f_{\theta^0_m}\right) \leq \frac{1}{2}\,\mathrm{e}^{\|\ell^0 - \ell^0_m\|_\infty} \| f^0 \|_\infty \| \ell^0 - \ell^0_m \|^2_{L^2} \leq \frac{1}{2}\,\mathrm{e}^{\gamma_m + \|\log(f^0)\|_\infty} \Delta_m^2.$$

□

Set:

$$\varepsilon_m = 6 d^{\frac{5}{2}} \kappa_m \Delta_m\, \mathrm{e}^{\left(4\gamma_m + 2\,\|\log(f^0)\|_\infty + 1\right)}. \tag{6.59}$$

We need the following lemma to control $\| \log(f^0/f_{\theta^*}) \|_\infty$.

Lemma 6.36. *If $\varepsilon_m \leq 1$, we also have:*

$$\| \log(f^0/f_{\theta^*}) \|_\infty \leq 2\gamma_m + \varepsilon_m \leq 2\gamma_m + 1. \tag{6.60}$$

Proof. To show (6.60), let $f^0_m = f_{\theta^0}$ denote the density function in the exponential family corresponding to θ^0, and $\alpha^0 = \int_\triangle \varphi_m f^0$. For each $1 \leq i \leq d$, the functions $\varphi_{i,m} = (\varphi_{[i],k}, 1 \leq k \leq m_i)$ form an orthonormal set with respect to the Lebesgue measure on $\triangle$. We set $\alpha^0_{i,m} = \int_\triangle \varphi_{i,m} f^0$ and $A_{i,m}(\theta^0) = \int_\triangle \varphi_{i,m} f_{\theta^0}$. By Bessel's inequality, we have for $1 \leq i \leq d$:

$$\| \alpha^0_{i,m} - A_{i,m}(\theta^0) \| \leq \| f^0 - f^0_m \|_{L^2}.$$

Summing up these inequalities for $1 \leq i \leq d$, we get:

$$\begin{aligned}
\| \alpha^0 - A_m(\theta^0) \| &\leq \sum_{i=1}^d \| \alpha^0_{i,m} - A_{i,m}(\theta^0) \| \\
&\leq d \| f^0 - f^0_m \|_{L^2} \\
&\leq d \| f^0 \|_\infty \mathrm{e}^{\left(\|\ell^0 - \ell^0_m\|_\infty - (\psi(\theta^0) - \mathrm{a}_0) \right)} \| \ell^0 - \ell^0_m \|_{L^2} \\
&\leq d \, \mathrm{e}^{\|\log(f^0)\|_\infty + 2\gamma_m} \, \Delta_m,
\end{aligned}$$

where we used (6.42) with $\kappa = \psi(\theta^0) - \mathrm{a}_0$ for the third inequality and $|\psi(\theta^0) - \mathrm{a}_0| \leq \gamma_m$ (due to $\psi(\theta^0) - \mathrm{a}_0 = \log(\int \exp(\ell^0_m - \ell^0) f^0)$) for the fourth inequality. The latter argument also ensures that $\| \log(f^0/f^0_m) \|_\infty \leq 2\gamma_m$. In order to apply Lemma 6.32 with $\theta = \theta^0$, $\alpha = \alpha^0$, we check condition (6.53), which is implied by:

$$d \, \mathrm{e}^{\|\log(f^0)\|_\infty + 2\gamma_m} \, \Delta_m \leq \frac{\mathrm{e}^{-(1 + |\log(f^0_m)\|_\infty)}}{6 d^{\frac{3}{2}} \kappa_m}.$$

Since $\| \log(f^0_m) \|_\infty \leq \| \log(f^0) \|_\infty + \| \log(f^0/f^0_m) \|_\infty \leq \| \log(f^0) \|_\infty + 2\gamma_m$, this condition is ensured whenever $\varepsilon_m \leq 1$. In this case, we deduce from (6.55) with $\tau = 1$ that $\| \log(f^0_m/f_{\theta^*}) \|_\infty \leq \varepsilon_m$. By the triangle inequality, we obtain $\| \log(f^0/f_{\theta^*}) \|_\infty \leq 2\gamma_m + \varepsilon_m$. This completes the proof. □

6.8.2 Variance of the estimator

We control the variance error due to the parameter estimation by the size of the sample. We keep the notations used in Section 6.8.1. In particular ε_m is defined by (6.59) and $\kappa_m = \sqrt{2d!} \sqrt{\sum_{i=1}^d (m_i + d)^{2d}}$. The results are summarized in the following proposition.

Proposition 6.37. *Let $f^0 \in \mathcal{P}(\triangle)$ have a product form given by Definition 6.1. Let $m \in (\mathbb{N}^*)^d$ and suppose that $\varepsilon_m \leq 1$. Set:*

$$\delta_{m,n} = 6 d^{\frac{3}{2}} \kappa_m \sqrt{\frac{|m|}{n}} \, \mathrm{e}^{2\gamma_m + \|\log(f^0)\|_\infty + 2}.$$

If $\delta_{m,n} \leq 1$, then for every $0 < K \leq \delta^{-2}_{m,n}$, we have:

$$\mathbb{P}\left(D\left(f_{\theta^*} \| \hat{f}_{m,n}\right) \geq \mathfrak{A}_2 \frac{|m|}{n} K \right) \leq \exp(\| \log(f^0) \|_\infty)/K. \tag{6.61}$$

where $\mathfrak{A}_2 = 3d \, \mathrm{e}^{2\gamma_m + \varepsilon_m + \|\log(f^0)\|_\infty + \tau}$, and $\tau = \delta_{m,n}\sqrt{K} \leq 1$.

Proof. Let θ^* be defined in Proposition 6.35. Let $X = (X_1, \ldots, X_d)$ denote a random variable with density f^0. Let θ in Lemma 6.32 be equal to θ^*, which gives $A_m(\theta^*) = \alpha^0 = \mathbb{E}[\varphi_m(X)]$, and for α, we take the empirical mean $\hat{\mu}_{m,n}$. With this setting, we have:

$$\| \alpha - \alpha^0 \|^2 = \sum_{i=1}^d \sum_{k=1}^{m_i} \left(\hat{\mu}_{m,n,i,k} - \mathbb{E}[\varphi_{i,k}(X_i)] \right)^2.$$

By Chebyshev's inequality $\|\alpha - \alpha^0\|^2 \leq |m| K/n$ except on a set whose probability verifies:

$$\mathbb{P}\left(\|\alpha - \alpha^0\|^2 > \frac{|m|}{n} K\right) \leq \frac{1}{|m| K} \sum_{i=1}^{d} \sum_{k=1}^{m_i} \sigma_{i,k}^2.$$

with $\sigma_{i,k}^2 = \mathrm{Var}\left[\varphi_{i,k}(X_i)\right]$. We have the upper bound $\sigma_{i,k}^2 \leq \|f^0\|_\infty \int_\triangle \varphi_{[i],k}^2 \leq \mathrm{e}^{\|\log(f^0)\|_\infty}$ by the normality of $\varphi_{i,k}$. Therefore we obtain:

$$\mathbb{P}\left(\|\alpha - \alpha^0\|^2 > \frac{|m|}{n} K\right) \leq \frac{\mathrm{e}^{\|\log(f^0)\|_\infty}}{K}.$$

We can apply Lemma 6.32 on the event $\{\|\alpha - \alpha^0\| \leq \sqrt{|m| K/n}\}$ if:

$$\sqrt{\frac{|m|}{n} K} \leq \frac{\mathrm{e}^{-(1+\|\log(f_{\theta^*})\|_\infty)}}{6 d^{\frac{3}{2}} \kappa_m}. \tag{6.62}$$

Thanks to (6.60) we have:

$$\|\log(f_{\theta^*})\|_\infty \leq \|\log(f^0/f_{\theta^*})\|_\infty + \|\log(f^0)\|_\infty \leq 2\gamma_m + \varepsilon_m + \|\log(f^0)\|_\infty. \tag{6.63}$$

Since $\varepsilon_m \leq 1$, (6.62) holds if $\delta_{m,n}^2 \leq 1/K$. Then except on a set of probability less than $\mathrm{e}^{\|\log(f^0)\|_\infty}/K$, the maximum likelihood estimator $\hat{\theta}_{m,n}$ satisfies, thanks to (6.56) with $\tau = \delta_{m,n}\sqrt{K}$:

$$D\left(f_{\theta^*} \| f_{\hat{\theta}_{m,n}}\right) \leq 3d\, \mathrm{e}^{\|\log(f_{\theta^*})\|_\infty + \tau} \frac{|m|}{n} K \leq 3d\, \mathrm{e}^{2\gamma_m + \varepsilon_m + \|\log(f^0)\|_\infty + \tau} \frac{|m|}{n} K. \tag{6.64}$$

□

6.8.3 Proof of Theorem 6.4

Recall that $r = (r_1, \ldots, r_d) \in \mathbb{N}^d$ is fixed. We assume $\ell_i^0 \in W_{r_i}^2(q_i)$ for all $1 \leq i \leq d$. Corollary 6.22 ensures $\Delta_m = O(\sqrt{\sum_{i=1}^d m_i^{-2r_i}})$ and the boundedness of γ_m when $m_i > r_i$ for all $1 \leq i \leq d$ is due to Corollary 6.24. By Remark 6.20, we have that $\kappa_m = O(|m|^d)$. If (6.10) holds, then $\kappa_m \Delta_m$ converges to 0. Therefore for m large enough, we have that ε_m defined in (6.59) is less than 1. By Proposition 6.35, the information projection f_{θ^*} of f^0 exists. For such m, by Lemma 6.28, we have that for all $\theta \in \mathbb{R}^{|m|}$:

$$D\left(f^0 \| f_\theta\right) = D\left(f^0 \| f_{\theta^*}\right) + D\left(f_{\theta^*} \| f_\theta\right).$$

Proposition 6.35 and $\Delta_m = O(\sqrt{\sum_{i=1}^d m_i^{-2r_i}})$ ensures that the $D\left(f^0 \| f_{\theta^*}\right) = O(\sum_{i=1}^d m_i^{-2r_i})$. The condition $\delta_{m,n} \leq 1$ in Proposition 6.37 is verified for n large enough since γ_m is bounded and (6.11) holds, giving $\lim_{n\to\infty} \delta_{m,n} = 0$. Proposition 6.37 then ensures that $D\left(f_{\theta^*} \| \hat{f}_{m,n}\right) = O_{\mathbb{P}}(|m|/n)$. Therefore the proof is complete.

6.9 Proof of Theorem 6.8

In this section we provide the elements of the proof of Theorem 6.8. We assume the hypotheses of Theorem 6.8. Recall the notation of Section 6.4. We shall stress out when we use the inequalities (6.20), (6.21) and (6.22) to achieve uniformity in r in Corollary 6.9.

First recall that ℓ^0 from (6.5) admits the following representation: $\ell^0 = \sum_{i=1}^d \sum_{k=1}^\infty \theta_{i,k}^0 \varphi_{[i],k}$. For $m = (m_1, \ldots, m_d) \in (\mathbb{N}^*)^d$, let $\ell_m^0 = \sum_{i=1}^d \sum_{k=1}^{m_i} \theta_{i,k}^0 \varphi_{[i],k}$ and $f_m^0 = \exp(\ell_m^0 - \psi(\theta_m^0))$. Using Corollary 6.24 and $|\psi(\theta_m^0) - \mathrm{a}_0| \leq \|\ell_m^0 - \ell^0\|_\infty$, we obtain that $\|\log(f_m^0/f^0)\|_\infty$ is bounded for all $m \in (\mathbb{N}^*)^d$ such that $m_i \geq r_i$:

$$\|\log(f_m^0/f^0)\|_\infty \leq 2\gamma_m \leq 2\gamma, \tag{6.65}$$

with $\gamma_m = \| \ell^0_m - \ell^0 \|_\infty$, and $\gamma = \mathcal{C} \sum_{i=1}^d \| \ell_i^{(r_i)} \|_{L^2(q_i)}$ with $\mathcal{C}$ defined in Corollary 6.24 which does not depend on r or m. For $m = (v, \dots, v) \in \mathcal{M}_n$, we have that $a_n \leq v \leq b_n$, with a_n, b_n given by:

$$a_n = \left\lfloor n^{1/(2(d+N_n)+1)} \right\rfloor \quad \text{and} \quad b_n = \left\lfloor n^{1/(2(d+1)+1)} \right\rfloor. \tag{6.66}$$

The upper bound (6.65) is uniform over $m \in \mathcal{M}_n$ and $r \in (\mathcal{R}_n)^d$ when (6.21) holds. Since $N_n = o(\log(n))$, we have $\lim_{n\to+\infty} a_n = +\infty$. Hence, for n large enough, say $n \geq n^*$, we have $\varepsilon_m \leq 1$ for all $m = (v, \dots, v) \in \mathcal{M}_n$ with ε_m given by (6.59), since $\kappa_m \Delta_m = O(a_n^{d-\min(r)})$. According to Proposition 6.35, this means that the information projection $f_{\theta^*_m}$ of f onto the set of functions $(\varphi_{[i],k}, 1 \leq i \leq d, 1 \leq k \leq v)$ verify, by (6.55) with $\tau = 1$, for all $m \in \mathcal{M}_n$:

$$\| \log(f_{\theta^*_m} / f^0_m) \|_\infty \leq 1. \tag{6.67}$$

Recall the notation $A^0_m = \int_\triangle \varphi_m f^0$ for the expected value of $\varphi_m(X^1)$, $\hat{\mu}_{m,n}$ the corresponding empirical mean based on the sample $\mathbb{X}_1^n$, and $\hat{\ell}_{m,n} = \hat{\theta}_{m,n} \cdot \varphi_m$ where $\hat{\theta}_{m,n}$ is the maximum likelihood estimate given by (6.9). Let $T_n > 0$ be defined as:

$$T_n = \frac{n_1 \, \mathrm{e}^{-4\gamma - 4 - 2\|\log(f^0)\|_\infty}}{36 d^5 d! b_n (b_n + d)^{2d} \log(b_n)}, \tag{6.68}$$

with b_n given by (6.66) and γ as in (6.65). We define the sets:

$$\mathcal{B}_{m,n} = \{ \| A^0_m - \hat{\mu}_{m,n} \|^2 > |m| \, T_n \log(b_n)/n_1 \} \quad \text{and} \quad \mathcal{A}_n = \left(\bigcup_{m \in \mathcal{M}_n} \mathcal{B}_{m,n} \right)^c.$$

We first show that with probability converging to 1, the estimators are uniformly bounded.

Lemma 6.38. *Let $n \in \mathbb{N}^*$, $n \geq n^*$ and $\mathcal{M}_n$ as in (6.14). Then we have:*

$$\mathbb{P}(\mathcal{A}_n) \geq 1 - N_n 2dn^{C_{T_n}},$$

with C_{T_n} defined as:

$$C_{T_n} = \frac{1}{2d+3} \left(1 - \frac{T_n}{2 \| f^0 \|_\infty + C\sqrt{T_n}} \right),$$

with a finite constant C given by (6.72). Moreover, on the event $\mathcal{A}_n$, we have the following uniform upper bound for $\| \hat{\ell}_{m,n} \|_\infty$, $m \in \mathcal{M}_n$:

$$\| \hat{\ell}_{m,n} \|_\infty \leq 4 + 4\gamma + 2 \| \log(f^0) \|_\infty. \tag{6.69}$$

Remark 6.39. Notice that by the definition of b_n, $\lim_{n\to\infty} T_n = +\infty$. For n large enough, we have $C_{T_n} < -\varepsilon < 0$ for some positive ε, so that:

$$\lim_{n\to\infty} N_n 2dn^{C_{T_n}} = 0. \tag{6.70}$$

This ensures that $\lim_{n\to\infty} \mathbb{P}(\mathcal{A}_n) = 1$, that is $(\hat{\ell}_{m,n}, m \in \mathcal{M}_n)$ are uniformly bounded with probability converging to 1.

Proof. For $m = (v, \dots, v) \in \mathcal{M}_n$ fixed, in order to bound the distance between the vectors $\hat{\mu}_{m,n} = (\hat{\mu}_{m,n,i,k}, 1 \leq i \leq d, 1 \leq k \leq v)$ and $A^0_m = \mathbb{E}[\hat{\mu}_{m,n}] = (\alpha^0_{i,k}, 1 \leq i \leq d, 1 \leq k \leq v)$, we first consider a single term $\left| \alpha^0_{i,k} - \hat{\mu}_{m,n,i,k} \right|$. By Bernstein's inequality, we have for all $t > 0$:

$$\begin{aligned}
\mathbb{P}\left(\left| \alpha^0_{i,k} - \hat{\mu}_{m,n,i,k} \right| > t \right) &\leq 2 \exp\left(- \frac{(n_1 t)^2/2}{n_1 \mathrm{Var}\, \varphi_{[i],k}(X^1) + 2 n_1 t \, \| \varphi_{i,k} \|_\infty / 3} \right) \\
&\leq 2 \exp\left(- \frac{(n_1 t)^2/2}{n_1 \mathbb{E}\left[\varphi^2_{[i],k}(X^1) \right] + 2 n_1 t \sqrt{2(d-1)!} (b_n + d)^{d - \frac{1}{2}} / 3} \right) \\
&\leq 2 \exp\left(- \frac{n_1 t^2/2}{\| f^0 \|_\infty + 2t \sqrt{2(d-1)!} (b_n + d)^{d - \frac{1}{2}} / 3} \right),
\end{aligned}$$

where we used, thanks to (6.32):

$$\| \varphi_{i,k} \|_\infty \leq \sqrt{(d-1)!}\sqrt{2k+d}\frac{(k+d-1)!}{k!} \leq \sqrt{2(d-1)!}(b_n+d)^{d-\frac{1}{2}}$$

for the second inequality, and the orthonormality of $\varphi_{[i],k}$ for the third inequality. Let us choose $t = \sqrt{T_n \log(b_n)/n_1}$. This gives:

$$\begin{aligned}\mathbb{P}\left(\left|\alpha^0_{i,k} - \hat{\mu}_{m,n,i,k}\right| > \sqrt{\frac{T_n \log(b_n)}{n_1}}\right) &\leq 2\exp\left(-\frac{T_n \log(b_n)/2}{\| f^0 \|_\infty + 2\sqrt{\frac{2T_n \log(b_n)(d-1)!(b_n+d)^{2d-1}}{9n_1}}}\right) \\ &\leq 2b_n^{-\frac{T_n}{2\|f^0\|_\infty + C\sqrt{T_n}}}, \end{aligned} \tag{6.71}$$

with C given by:

$$C = \sup_{n\in\mathbb{N}^*} 4\sqrt{\frac{2\log(b_n)(d-1)!(b_n+d)^{2d-1}}{9n_1}}. \tag{6.72}$$

Notice $C < +\infty$ since the sequence $\sqrt{\log(b_n)(b_n+d)^{2d-1}/9n_1}$ is $o(1)$. For the probability of $\mathcal{B}_{n,m}$ we have:

$$\begin{aligned}\mathbb{P}\left(\mathcal{B}_{n,m}\right) &\leq \sum_{i=1}^{d}\sum_{k=1}^{v} \mathbb{P}\left(\left|\alpha^0_{i,k} - \hat{\mu}_{m,n,i,k}\right|^2 > \frac{T_n \log(b_n)}{n_1}\right) \\ &\leq \sum_{i=1}^{d}\sum_{k=1}^{v} 2b_n^{-\frac{T_n}{2\|f^0\|_\infty + C\sqrt{T_n}}} \\ &\leq 2dn^{C_{T_n}}.\end{aligned}$$

This implies the following lower bound on $\mathbb{P}(\mathcal{A}_n)$:

$$\mathbb{P}(\mathcal{A}_n) = 1 - \mathbb{P}\left(\bigcup_{m\in\mathcal{M}_n} \mathcal{B}_{n,m}\right) \geq 1 - \sum_{m\in\mathcal{M}_n} \mathbb{P}(\mathcal{B}_{n,m}) \geq 1 - N_n 2dn^{c_{T_n}}.$$

On $\mathcal{A}_n$, by the definition of T_n, we have for all $m \in \mathcal{M}_n$:

$$\| A^0_m - \hat{\mu}_{m,n} \| 6d^2\sqrt{2d!}(v+d)^d \, \mathrm{e}^{2\gamma_m+2} \leq \sqrt{b_n \frac{T_n \log(b_n)}{n_1}} 6d^{\frac{5}{2}}\sqrt{2d!}(b_n+d)^d \, \mathrm{e}^{2\gamma+2} = 1.$$

Notice that whenever (6.9) holds, condition (6.53) of Lemma 6.32 is satisfied with $\theta = \theta^*_m$ and $\alpha = \hat{\mu}_{m,n}$, thanks to $\kappa_m \leq \sqrt{d2d!}(b_n+d)^d$ and:

$$\| \log(f_{\theta^*_m}) \|_\infty \leq \| \log(f_{\theta^*_m}/f^0_m) \|_\infty + \| \log(f^0_m/f^0) \|_\infty + \| \log(f^0) \|_\infty \leq 1 + 2\gamma + \| \log(f^0) \|_\infty.$$

According to Equation (6.55) with $\tau = 1$, we can deduce that on $\mathcal{A}_n$, we have:

$$\| \log(\hat{f}_{m,n}/f_{\theta^*_m}) \|_\infty \leq 1 \quad \text{for all } m \in \mathcal{M}_n, n \geq n^*.$$

This, along with (6.65) and (6.67), provide the following uniform upper bound for $(\| \hat{\ell}_{m,n} \|_\infty, m \in \mathcal{M}_n)$ on $\mathcal{A}_n$:

$$\begin{aligned}\frac{1}{2}\| \hat{\ell}_{m,n} \|_\infty &\leq \| \log(\hat{f}_{m,n}) \|_\infty \\ &\leq \| \log(\hat{f}_{m,n}/f_{\theta^*_m}) \|_\infty + \| \log(f_{\theta^*_m}/f^0_m) \|_\infty + \| \log(f^0_m/f^0) \|_\infty + \| \log(f^0) \|_\infty \\ &\leq 2 + 2\gamma + \| \log(f^0) \|_\infty,\end{aligned}$$

where we used (6.44) for the first inequality.

□

We also give a sharp oracle inequality for the convex aggregate estimator $f_{\hat{\lambda}_n^*}$ conditionally on $\mathcal{A}_n$ with n fixed . The following lemma is a direct application of Theorem 3.6. of [38] and (6.69).

Lemma 6.40. *Let $n \in \mathbb{N}^*$ be fixed. Conditionally on $\mathcal{A}_n$, let $f_{\hat{\lambda}_n^*}$ be given by (6.16) with $\hat{\lambda}_n^*$ defined as in (6.18). Then for any $x > 0$ we have with probability greater than $1 - \exp(-x)$:*

$$D\left(f^0 \| f_{\hat{\lambda}_n^*}\right) - \min_{m \in \mathcal{M}_n} D\left(f^0 \| \hat{f}_{m,n}\right) \leq \frac{\beta(\log(N_n) + x)}{n_2}, \tag{6.73}$$

with $\beta = 2\exp(6K + 2L) + 4K/3$, and $L, K \in \mathbb{R}$ given by :

$$L = \| \ell^0 \|_\infty, \qquad K = 4 + 4\gamma + 2 \| \log(f^0) \|_\infty,$$

with γ as in (6.65).

Now we prove Theorem 6.8. For $n \in \mathbb{N}^*$ and $C > 0$, we define the event $\mathcal{D}_n(C)$ as:

$$\mathcal{D}_n(C) = \left\{ D\left(f^0 \| f_{\hat{\lambda}_n^*}\right) \geq C \left(n^{-\frac{2\min(r)}{2\min(r)+1}}\right) \right\}.$$

Let $\varepsilon > 0$. To prove (6.19), we need to find $C_\varepsilon > 0$ such that for all n large enough:

$$\mathbb{P}\left(\mathcal{D}_n(C_\varepsilon)\right) \leq \varepsilon. \tag{6.74}$$

We decompose the left hand side of (6.74) according to $\mathcal{A}_n$:

$$\mathbb{P}\left(\mathcal{D}_n(C_\varepsilon)\right) \leq \mathbb{P}\left(\mathcal{D}_n(C_\varepsilon) \,|\, \mathcal{A}_n\right) \mathbb{P}(\mathcal{A}_n) + \mathbb{P}(\mathcal{A}_n^c). \tag{6.75}$$

The product $\mathbb{P}\left(\mathcal{D}_n(C_\varepsilon) \,|\, \mathcal{A}_n\right) \mathbb{P}(\mathcal{A}_n)$ is bounded by:

$$\mathbb{P}\left(\mathcal{D}_n(C_\varepsilon) \,|\, \mathcal{A}_n\right) \mathbb{P}(\mathcal{A}_n) \leq A_n(C_\varepsilon) + B_n(C_\varepsilon),$$

with $A_n(C_\varepsilon)$ and $B_n(C_\varepsilon)$ defined by:

$$A_n(C_\varepsilon) = \mathbb{P}\left(D\left(f^0 \| f_{\hat{\lambda}_n^*}\right) - \min_{m \in \mathcal{M}_n} D\left(f^0 \| \hat{f}_{m,n}\right) \geq \frac{C_\varepsilon}{2} \left(n^{-\frac{2\min(r)}{2\min(r)+1}}\right) \,\middle|\, \mathcal{A}_n \right),$$

$$B_n(C_\varepsilon) = \mathbb{P}\left(\min_{m \in \mathcal{M}_n} D\left(f^0 \| \hat{f}_{m,n}\right) \geq \frac{C_\varepsilon}{2} \left(n^{-\frac{2\min(r)}{2\min(r)+1}}\right) \right).$$

To bound $A_n(C_\varepsilon)$ we apply Lemma 6.40 with $x = x_\varepsilon = -\log(\varepsilon/4)$:

$$\mathbb{P}\left(D\left(f^0 \| f_{\hat{\lambda}_n^*}\right) - \min_{m \in \mathcal{M}_n} D\left(f^0 \| \hat{f}_{m,n}\right) \geq \frac{\beta(\log(N_n) + x_\varepsilon)}{n_2} \,\middle|\, \mathcal{A}_n \right) \leq \frac{\varepsilon}{4}.$$

Let us define $C_{\varepsilon,1}$ as:

$$C_{\varepsilon,1} = \sup_{n \in \mathbb{N}^*} \left(\frac{\beta(\log(N_n) + x_\varepsilon)}{n_2 n^{-\frac{2\min(r)}{2\min(r)+1}}} \right). \tag{6.76}$$

Since $N_n = o(\log(n))$, we have $C_{\varepsilon,1} < +\infty$ as the sequence on the right hand side of (6.76) is $o(1)$. This bound is uniform over regularities in $(\mathcal{R}_n)^d$ thanks to (6.22) Therefore for all $C_\varepsilon \geq C_{\varepsilon,1}$, we have $A_n(C_\varepsilon) \leq \varepsilon/4$.

For $B_n(C_\varepsilon)$, notice that if $n \geq \bar{n}$ with $\bar{n}$ given by (6.15), then $m^* = (v^*, \ldots, v^*) \in \mathcal{M}_n$ with $v^* = \lfloor n^{1/(2\min(r)+1)} \rfloor$. This holds for all $r \in (\mathcal{R}_n)^d$ due to (6.20). By Remark 6.5, we have that $D\left(f^0 \| \hat{f}_{m^*,n}\right) = O_{\mathbb{P}}(n^{-2\min(r)/(2\min(r)+1)})$. This ensure that there exists $C_{\varepsilon,2}$ such that for all $C_\varepsilon \geq C_{\varepsilon,2}$, $n \geq \bar{n}$:

$$B_n(C_\varepsilon) \leq \mathbb{P}\left(D\left(f^0 \| \hat{f}_{m^*,n}\right) \geq \frac{C_{\varepsilon,2}}{2} \left(n^{-\frac{2\min(r)}{2\min(r)+1}}\right) \right) \leq \frac{\varepsilon}{4}.$$

We also have by (6.70) that there exists $\tilde{n} \in \mathbb{N}^*$ such that $\mathbb{P}(\mathcal{A}_n^c) \leq \varepsilon/2$ for all $n \geq \tilde{n}$. Therefore by setting $C_\varepsilon = \max(C_{\varepsilon,1}, C_{\varepsilon,2})$ in (6.75), we have for all $n \geq \max(n^*, \bar{n}, \tilde{n})$:

$$\mathbb{P}\left(\mathcal{D}_n(C_\varepsilon)\right) \leq A_n(C_\varepsilon) + B_n(C_\varepsilon) + \mathbb{P}(\mathcal{A}_n^c) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

which gives (6.74) and thus concludes the proof.

Chapter 7

Application to nuclear engineering data

7.1 Industrial context and the dataset

In this section, we apply the proposed methodology to estimate the joint distribution of the dimensions of flaws of a passive component in an EDF electric power plant. These flaws may lead to a crack under the severe stress to which the material is exposed, endangering the integrity of the component. The model predicting the propagation of the flaws requires its size (given by Length $\times$ Depth) as an input parameter, therefore the joint modelling of the distribution of these two quantities is crucial. Since higher values of the size of the flaws are more penalizing for the occurrence of a crack, we prefer a model which is not only adequate for the dataset, but assigns relatively great probability to higher values of these dimensions to obtain a conservative estimation of the failure probability of the component.

EDF possesses a database of joint measurements of these quantities which contains $n = 198$ measurements obtained by supervised experimentations along with 341 observations registered during regular inspections of the components in operation. We will only consider the database coming from the experimentations as these can be considered statistically perfect, whereas the inspection data is subject to measurement uncertainty and detection threshold.

Both sets of data suggest that the dimensions verify the ordering constraint, since for every pair of dimensions we have that the length of the flaw is greater than the depth. The currently applied modelling schemes does not take into consideration this aspect of the dataset. Figure 7.1 presents the experimentation dataset after applying a strictly monotone transformation on both dimensions to obtain values on $[0, 1]$.

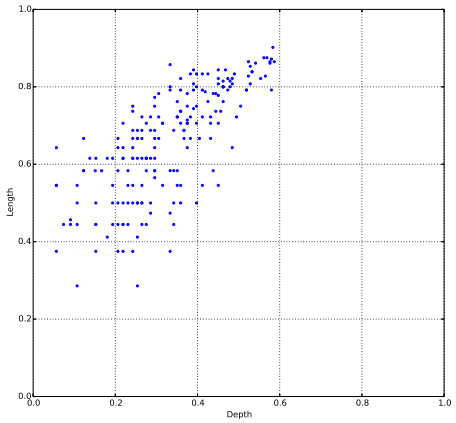


Figure 7.1 – Scatter-plot of the transformed data set

7.2 Available modelling schemes

We will compare our approach to the currently approved modelling scheme as well as a method proposed by the former conference paper [148]. In what follows, we note by L the random variable of the length of the flaw and by D the random variable of the depth.

7.2.1 Reference model

The first method used in current statistical studies of this problem at EDF consists of modelling the joint distribution of the pair (D, R), where $R = D/L$ is the random variable of the ratio between the two dimensions. The model takes the assumption that D and R are independent, and propose the following distributions for these variables (we omit the parameters of the distributions for confidentiality reasons):

- F_D: Weibull with two parameters,
- F_R: Log-normal.

7.2.2 Parametric model

[148] proposes a parametric copula-based approach for modelling the dependence structure between D and L. Copula theory allows to separate the modelling of the marginals and the dependence structure. The joint distribution function $F_{(D,L)}$ of the pair (D, L) can be expressed by Sklar's Theorem as:

$$F_{(D,L)}(d, l) = C_{(D,L)}(F_D(d), F_L(l)),$$

where F_D, F_L are the marginal distribution functions of D and L, and $C_{(D,L)}$ is the connecting copula containing all information on the dependence. We refer to [132] for an overview of copula theory. In this setting, both dimensions D and L are modelled by Weibull distributions with two parameters. For the connecting copula $C_{(D,L)}$ the Authors consider multiple parametric families such as Gaussian, Frank or Gumbel copulas. They estimate, based on the dataset, the parameters in each family by various methods, and compare the resulting joint distributions in order to determine the most relevant model. In conclusion, the Gumbel copula proved to give the most satisfactory results according to the graphical criterion of the Kendall plots and the Cramér-von-Mises goodness-of-fit test.

7.3 Estimation of the nonparametric model

For the nonparametric model, we have first transformed the dataset by using the monotone transformation T given by, for $x \in \mathbb{R}^+$:

$$T(x) = \frac{cx}{cx + 1},$$

with c a constant. This is necessary since the estimation procedure requires a sample distributed on $\triangle$. The impact of the choice of the transformation function T as well as the constant c on the estimation quality has not been addressed in this paper. We choose an equal number of parameters $m = m_1 = m_2$ for both dimensions. We estimate the parameters $\theta = (\theta_{i,k}; 1 \leq i \leq 2, 1 \leq k \leq m)$ by maximizing the function G given by:

$$G(\theta) = \sum_{i=1}^{2} \sum_{k=1}^{m} \theta_{i,k} \hat{\mu}_{i,k} - \psi(\theta)$$

with $\hat{\mu}_{1,k} = (1/n) \sum_{j=1}^{n} \varphi_{1,k}(D^j)$, $\hat{\mu}_{2,k} = (1/n) \sum_{j=1}^{n} \varphi_{2,k}(L^j)$, and $\psi(\theta)$ is given by $\psi(\theta) = \log(\int_{\triangle} \exp(\sum_{i=1}^{2} \sum_{k=1}^{m} \theta_{i,k} \varphi_{i,k}(x_i))\, dx)$. This is equivalent to solving equation (6.9). We estimate our model for increasing values of m, using the result of the previous estimation with fewer parameters as described in [176]. We use the TNC algorithm of the OpenTURNS library for Python to numerically maximize G. The estimated parameters for $m = 1, 2, 3, 4$ can be found in Table 7.1.

Table 7.1 – Estimated parameters for $m = 1, 2, 3, 4$.

m	$\hat{\theta}_{1,k}$	$\hat{\theta}_{2,k}$
1	$\hat{\theta}_{1,1} = -0.000307772$	$\hat{\theta}_{2,1} = -0.0277476$
2	$\hat{\theta}_{1,1} = -0.523519$	$\hat{\theta}_{2,1} = 0.295835$
	$\hat{\theta}_{1,2} = -1.06206$	$\hat{\theta}_{2,2} = -0.814702$
3	$\hat{\theta}_{1,1} = -0.545568$	$\hat{\theta}_{2,1} = -0.0445993$
	$\hat{\theta}_{1,2} = -1.10107$	$\hat{\theta}_{2,2} = -0.603401$
	$\hat{\theta}_{1,3} = -0.00991902$	$\hat{\theta}_{2,3} = -0.310838$
4	$\hat{\theta}_{1,1} = -1.82941$	$\hat{\theta}_{2,1} = 0.759716$
	$\hat{\theta}_{1,2} = -2.73921$	$\hat{\theta}_{2,2} = -2.43278$
	$\hat{\theta}_{1,3} = -1.1029$	$\hat{\theta}_{2,3} = 0.626079$
	$\hat{\theta}_{1,4} = -0.631885$	$\hat{\theta}_{2,4} = -1.03101$

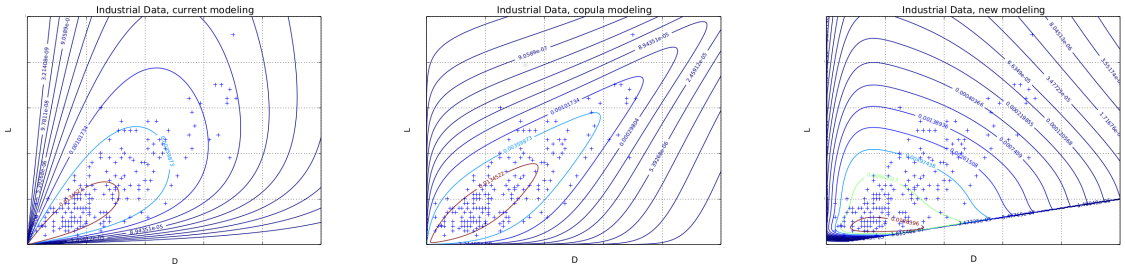
Table 7.2 – Log-likelihood and BIC of the competing models with empirical data.

Model	Copula	Log-likelihood	BIC
Reference	-	-957.399	1930.663
Parametric	Gumbel	-927.196	1880.833
Nonparametric	MaxEntropy	-998.516	2039.338

7.4 Comparison of the competing models

7.4.1 Fitting to the empirical data

Here we compare the three different approaches in terms of goodness-of-fit to the underlying dataset and the resulting failure probability. For the reference and parametric model, we utilize the parameters obtained in the previous studies. For the nonparametric model, we take $m_1 = m_2 = 4$. In Figure 7.2, the densities obtained from each model can be seen along with the dataset. One can observe that the support of the nonparametric model is indeed the half plane S, whereas the other two models allow the variables to take values such that $L < D$. In Table 7.2 we calculated the log-likelihood of each model along with the BIC value. According to these values, the parametric model seems the most adapted for the sample followed by the reference model and the nonparametric model. The results suggest that distribution of the sample may not belong to the family of maximum entropy distributions of order statistics, and there may exist a hidden constraint that needs to be taken into consideration.



(a) Reference model (b) Parametric model (Gumbel) (c) Nonparametric model

Figure 7.2 – Isodensities for the competing models with empirical data.

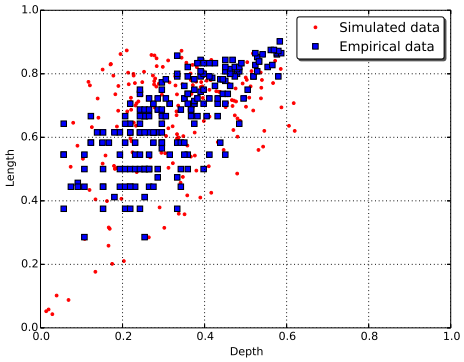


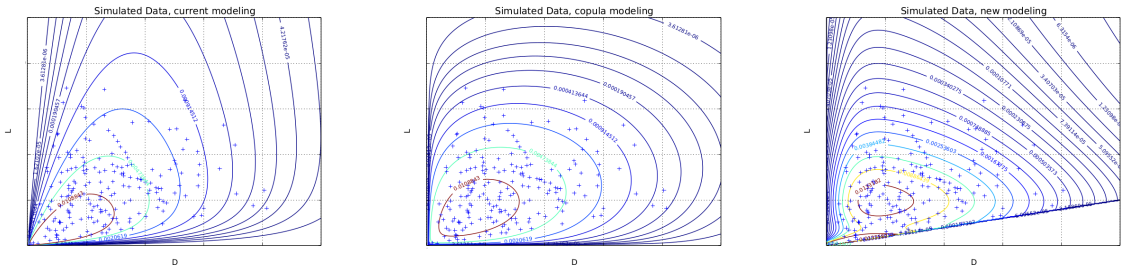
Figure 7.3 – Scatter-plot of the two transformed data sets.

Table 7.3 – Log-likelihood and BIC of the competing models with simulated data.

Model	Copula	Log-likelihood	BIC
Reference	-	−1050.075	2116.016
Parametric	Frank	−1031.315	2089.072
Parametric	Gumbel	−1030.492	2087.425
Parametric	Normal	−1021.243	2068.928
Nonparametric	MaxEntropy	−995.058	2032.423

7.4.2 Fitting to simulated data

In order to show that effectiveness of the nonparametric model when the underlying distribution belongs to the family of maximum entropy distributions of order statistics, we simulate a dataset with 198 entries from the maximum entropy distribution with the same Weibull marginals which were used to construct the parametric model in 7.2.2. Figure 7.3 shows the difference between the two sets of data. We re-estimated all the parameters of the three competing models, and Table 7.3 shows the log-likelihood and BIC values for each model. For the parametric model, we made estimations using the Frank, Gumbel and Normal (Gaussian) family of copulas. The results confirm that if the underlying distribution belongs to the family of maximum entropy distributions of order statistics, then the nonparametric model outperforms the reference and parametric models.



(a) Reference model (b) Parametric model (Normal) (c) Nonparametric model

Figure 7.4 – Isodensities for the competing models with simulated data.

7.4.3 Failure probability

We use the joint distribution of the pair (D, L) estimated in three different ways in 7.4.1 to carry out a Monte Carlo study to determine the impact of the modelling on the component

Table 7.4 – Failure probabilities calculated with the competing models using an importance sampling method with 10^{-4} simulations.

Model	$\hat{R}^f_{model}$	c_{model}
Reference	1	2.09%
Parametric	1.022	1.99%
Nonparametric	0.148	4.14%

failure probability. The failure probability P^f is the probability that one of the output factors of the fracture mechanics model stays below a certain threshold. To estimate this probability, we couple the fracture mechanics model with the OpenTURNS platform. The fracture mechanics model takes 15 input variables, we assume that the pair (L, D) is independent of the rest of the variables whose values are fixed at an average level for this study. We evaluate the failure probabilities by Monte-Carlo simulations with importance sampling using $N = 10^4$ simulations. The simulations provide the estimators $\hat{P}^f_{model}$ for the three models. The results are summarized in Table 7.4, where we give the estimated failure probabilities relative to the failure probability of the reference model, that is the ratio:

$$\hat{R}^f_{model} = \frac{\hat{P}^f_{model}}{\hat{P}^f_{ref.model}},$$

and the coefficient of variation c_{model} given by:

$$c_{model} = \sqrt{\frac{1 - \hat{P}^f_{model}}{\hat{P}^f_{model} N}}.$$

We observe that the nonparametric model estimates the failure probability to be much lower than the other two models. This is due to the fact that a failure usually occurs when both D and L assume high values. The Gumbel copula ensures a high positive tail dependence, leading to more frequent common high values, whereas the nonparametric model, as Figure 7.2 suggests, gives more probabilistic mass to the upper-left zones with greater L values but smaller D values.

7.5 Conclusions

In this paper we draw attention to the importance of modelling the dependence structure of random variables appearing in uncertainty quantification studies. The modelling should take into consideration all the available statistical data, but ensure a maximum of freedom besides this knowledge. We presented the family of maximum entropy distribution of ordered random variables as well as a nonparametric estimation procedure to efficiently estimate such distributions. We examined its statistical performance in an uncertainty quantification study compared to some other approaches. We have seen that when the underlying data set comes from a distribution which belongs to the family of maximum entropy distributions of order statistics, the nonparametric density estimation approach proposed in Section 6 performs well. When applied to the industrial case study, we observe a decline in the performance of the nonparametric estimator, suggesting that there are some hidden constraints in addition to the ordering which was not taken into consideration by this approach (for example the high upper tail dependence). The failure probability calculations shows that the dependence modelling have a significant impact on the estimation of failure risks.

In following studies we would like to determine, via extensive simulation studies, the cases where such distributions may give more favourable results compared to other approaches. We would like to give a testing procedure to determine whether the underlying data set comes

from a maximum entropy distribution or there are other hidden constraints which need to be taken into consideration. An aggregation method is also under development to give an adaptive nonparametric estimator of the maximum entropy distribution which performs as well as the nonparametric model with an optimal number of parameters which depends on the density's unknown regularity.

Bibliography

[1] R. Abraham, J.-F. Delmas, and H. Guo. Critical multi-type galton-watson trees conditioned to be large. *arXiv preprint arXiv:1511.01721*, 2015.

[2] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Courier Dover Publications, 1970.

[3] I. S. Abramson. On bandwidth variation in kernel estimates-a square root law. *The Annals of Statistics*, 10(4):1217–1223, 1982. ISSN 00905364. URL http://www.jstor.org/stable/2240724.

[4] A. AghaKouchak. Entropy–copula in hydrology and climatology. *Journal of Hydrometeorology*, 15(6):2176–2189, 2014.

[5] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja. *A first course in order statistics*, volume 54. Siam, 1992.

[6] P. Asghari, V. Fakoor, and M. Sarmad. A berry-esseen type bound in kernel density estimation for a random left-truncation model. *Communications for Statistical Applications and Methods*, 21(2):115–124, April 2014. ISSN 2287-7843.

[7] O. Asserin, A. Loredo, M. Petelet, and B. Iooss. Global sensitivity analysis in welding simulations - what are the material data you really need? *Finite Elements in Analysis and Design*, 47(9):1004–1016, 2011.

[8] J.-Y. Audibert. Progressive mixture rules are deviation suboptimal. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 41–48. Curran Associates, Inc., 2008.

[9] J. Avérous, C. Genest, and S. C. Kochar. On the dependence structure of order statistics. *Journal of Multivariate Analysis*, 94(1):159–171, 2005.

[10] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

[11] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *Information Theory, IEEE Transactions on*, 37(4):1034–1054, 1991.

[12] A. R. Barron and C.-H. Sheu. Approximation of density functions by sequences of exponential families. *The Annals of Statistics*, 19(3):1347–1369, 1991.

[13] A. R. Barron, L. Gyorfi, and E. C. van der Meulen. Distribution estimation consistent in total variation and in two types of information divergence. *Information Theory, IEEE Transactions on*, 38(5):1437–1454, 1992.

[14] M. S. Bartlett. Statistical estimation of density functions. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 245–254, 1963.

[15] M. Baudin, A. Dutfoy, B. Iooss, and A.-L. Popelin. Open TURNS: An industrial software for uncertainty quantification in simulation. *arXiv preprint arXiv:1501.05242*, 2015.

[16] T. Bedford and K. Wilson. On the construction of minimum information bivariate copula families. *Annals of the Institute of Statistical Mathematics*, pages 1–21, 2013. ISSN 0020-3157. doi: 10.1007/s10463-013-0422-0. URL http://dx.doi.org/10.1007/s10463-013-0422-0.

[17] P. Bellec. Optimal exponential bounds for aggregation of density estimators. *arXiv preprint arXiv:1405.3907*, 2014.

[18] O. Benrabah, E. Ould Saïd, and A. Tatachak. A kernel mode estimate under random left truncation and time series model: asymptotic normality. *Statistical Papers*, 56(3):887–910, 2014. ISSN 1613-9798. doi: 10.1007/s00362-014-0613-7. URL http://dx.doi.org/10.1007/s00362-014-0613-7.

[19] K. Bertin. Asymptotically exact minimax estimation in sup-norm for anisotropic Hölder classes. *Bernoulli*, 10(5):873–888, 2004.

[20] S. Bertino. Sulla dissomiglianza tra mutabili cicliche. *Metron*, 35:53 – 88, 1977.

[21] P. J. Bickel. Some contributions to the theory of order statistics. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calf., 1965/66), Vol. I: Statistics*, pages 575–591. Univ. California Press, Berkeley, Calif., 1967.

[22] J. Bigot, R. B. Lirio, J.-M. Loubes, and L. M. Alvarez. Adaptive estimation of spectral densities via wavelet thresholding and information projection. *arXiv preprint arXiv:0912.2026*, 2009.

[23] L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam*, pages 55–87. Springer New York, 1997. ISBN 978-1-4612-7323-3. doi: 10.1007/978-1-4612-1880-7_4. URL http://dx.doi.org/10.1007/978-1-4612-1880-7_4.

[24] P. J. Boland, M. Hollander, K. Joag-Dev, and S. Kochar. Bivariate dependence properties of order statistics. *Journal of Multivariate Analysis*, 56(1):75–89, 1996.

[25] J. Borwein, A. Lewis, and R. Nussbaum. Entropy minimization, *DAD* problems, and doubly stochastic kernels. *Journal of Functional Analysis*, 123(2):264 – 307, 1994. ISSN 0022-1236. doi: http://dx.doi.org/10.1006/jfan.1994.1089. URL http://www.sciencedirect.com/science/article/pii/S0022123684710895.

[26] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[27] R. C. Bradley. On positive spectral density functions. *Bernoulli*, 8(2):175–193, 2002.

[28] L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19(2):135–144, 1977.

[29] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 08 2007. doi: 10.1214/009053606000001587. URL http://dx.doi.org/10.1214/009053606000001587.

[30] C. Butucea. Exact adaptive pointwise estimation on sobolev classes of densities. *ESAIM: Probability and Statistics*, 5:1–31, 2001.

[31] C. Butucea, J. Delmas, A. Dutfoy, and R. Fischer. Modélisation de la dépendance sous contrainte déterministe. In *Proceedings of Congrès Lambda Mu 19 de Maîtrise des Risques et Sûreté de Fonctionnement*, 2014. doi: 10.4267/2042/56165.

[32] C. Butucea, J.-F. Delmas, A. Dutfoy, and R. Fischer. Copule d'entropie maximale avec des statistiques d'ordre fixées. In *Proceedings of Journées de Statistiques Rennes*, 2014. URL http://papersjds14.sfds.asso.fr/submission_200.pdf.

[33] C. Butucea, J.-F. Delmas, A. Dutfoy, and R. Fischer. Maximum entropy copula with given diagonal section. *Journal of Multivariate Analysis*, 137:61 – 81, 2015. ISSN 0047-259X. doi: http://dx.doi.org/10.1016/j.jmva.2015.01.003. URL http://www.sciencedirect.com/science/article/pii/S0047259X15000081.

[34] C. Butucea, J.-F. Delmas, A. Dutfoy, and R. Fischer. Nonparametric estimation of distributions of order statistics with application to nuclear engineering. In L. Podofillini, B. Sudret, B. Stojadinovic, E. Zio, and W. Kröger, editors, *Safety and Reliability of Complex Engineered Systems: ESREL 2015*. CRC Press, 2015. ISBN 9781315648415. URL https://books.google.fr/books?id=C9GYCgAAQBAJ.

[35] C. Butucea, J.-F. Delmas, A. Dutfoy, and R. Fischer. Estimation rapide non-paramétrique de la densité de la distribution d'entropie maximale pour les statistiques d'ordre. In *Proceedings of Journées de Statistiques Rennes*, 2015. URL http://papersjds15.sfds.asso.fr/submission_83.pdf.

[36] C. Butucea, J.-F. Delmas, A. Dutfoy, and R. Fischer. Maximum entropy distribution of order statistics with given marginals. *arXiv preprint arXiv:1509.02019*, 2015.

[37] C. Butucea, J.-F. Delmas, A. Dutfoy, and R. Fischer. Fast adaptive estimation of log-additive exponential models in kullback-leibler divergence. *arXiv preprint arXiv:1604.06304*, 2016.

[38] C. Butucea, J.-F. Delmas, A. Dutfoy, and R. Fischer. Optimal exponential bounds for aggregation of estimators for the kullback-leibler loss. *arXiv preprint arXiv:1601.05686*, 2016.

[39] T. Cacoullos. Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 18(1):179–189, 1966.

[40] O. Catoni. The mixture approach to universal model selection. In *École Normale Supérieure*. Citeseer, 1997.

[41] O. Catoni. Universal aggregation rules with exact bias bounds. Laboratoire de Probabilités et Modeles Aléatoires, CNRS, Paris. *Preprint*, 510, 1999.

[42] N. Cencov. Estimation of an unknown density function from observations. In *Dokl. Akad. Nauk, SSSR*, volume 147, pages 45–48, 1962.

[43] C. Chang and D. Politis. Aggregation of spectral density estimators. *Statistics & Probability Letters*, 94:204–213, 2014.

[44] E. Chicken and T. T. Cai. Block thresholding for density estimation: local and global adaptivity. *Journal of Multivariate Analysis*, 95(1):76 – 106, 2005. ISSN 0047-259X. doi: http://dx.doi.org/10.1016/j.jmva.2004.07.003. URL http://www.sciencedirect.com/science/article/pii/S0047259X04001563.

[45] Y.-S. Chow, S. Geman, and L.-D. Wu. Consistent cross-validated density estimation. *The Annals of Statistics*, pages 25–38, 1983.

[46] B. Chu. Recovering copulas from limited information and an application to asset allocation. *Journal of Banking & Finance*, 35(7):1824–1842, 2011.

[47] B. R. Crain. An information theoretic approach to approximating a probability distribution. *SIAM Journal on Applied Mathematics*, 32(2):339–346, 1977.

[48] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probability*, 3:146–158, 1975.

[49] I. Cuculescu and R. Theodorescu. Copulas: diagonals, tracks. *Revue roumaine de mathématiques pures et appliquées*, 46(6):731–742, 2001.

[50] D. Dai, P. Rigollet, and T. Zhang. Deviation optimal learning using greedy Q-aggregation. *Ann. Statist.*, 40(3):1878–1905, 06 2012. doi: 10.1214/12-AOS1025. URL http://dx.doi.org/10.1214/12-AOS1025.

[51] D. Dai, P. Rigollet, L. Xia, and T. Zhang. Aggregation of affine estimators. *Electron. J. Statist.*, 8(1):302–327, 2014. doi: 10.1214/14-EJS886. URL http://dx.doi.org/10.1214/14-EJS886.

[52] A. S. Dalalyan and J. Salmon. Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.*, 40(4):2327–2355, 08 2012. doi: 10.1214/12-AOS1038. URL http://dx.doi.org/10.1214/12-AOS1038.

[53] A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory*, pages 97–111. Springer, 2007.

[54] A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61, 2008.

[55] H. A. David and H. N. Nagaraja. *Order statistics*. Wiley Online Library, 1970.

[56] R. B. Davies. Asymptotic inference in stationary Gaussian time-series. *Advances in Appl. Probability*, 5:469–497, 1973. ISSN 0001-8678.

[57] E. de Amo, M. D. Carrillo, and J. F. Sánchez. Absolutely continuous copulas with given sub-diagonal section. *Fuzzy Sets and Systems*, 228(0):105 – 113, 2013. ISSN 0165-0114. doi: http://dx.doi.org/10.1016/j.fss.2012.10.002. URL http://www.sciencedirect.com/science/article/pii/S0165011412004290. Special issue on *AGOP* 2011 and EUSFLAT/LFA 2011.

[58] E. de Amo, H. D. Meyer, M. D. Carrillo, and J. F. Sánchez. Characterization of copulas with given diagonal and opposite diagonal sections. *Fuzzy Sets and Systems*, 284: 63 – 77, 2016. ISSN 0165-0114. doi: http://dx.doi.org/10.1016/j.fss.2014.10.030. URL http://www.sciencedirect.com/science/article/pii/S0165011414004862. Theme: Uncertainty and Copulas.

[59] B. V. de Melo Mendes and M. A. Sanfins. The limiting copula of the two largest order statistics of independent and identically distributed samples. *Brazilian Journal of Probability and Statistics*, 21:85–101, 2007.

[60] G. F. de Montricher, R. A. Tapia, and J. R. Thompson. Nonparametric maximum likelihood estimation of probability densities by penalty function methods. *The Annals of Statistics*, pages 1329–1348, 1975.

[61] M. A. Dempster, E. A. Medova, and S. W. Yang. Empirical copulas for CDO tranche pricing using relative entropy. *International Journal of Theoretical and Applied Finance*, 10(04):679–701, 2007.

[62] L. Depradeux. *Simulation numérique du soudage-acier 316L: validation sur cas tests de complexité croissante*. PhD thesis, Villeurbanne, INSA, 2004.

[63] D. Devroye, J. Beirlant, R. Cao, R. Fraiman, P. Hall, M. Jones, G. Lugosi, E. Mammen, J. Marron, C. Sánchez-Sellero, et al. Universal smoothing factor selection in density estimation: theory and practice. *Test*, 6(2):223–320, 1997.

[64] E. Di Nezza, G. Palatucci, and E. Valdinoci. Hitchhiker's guide to the fractional Sobolev spaces. *Bull. Sci. Math.*, 136(5):521–573, 2012. ISSN 0007-4497. doi: 10.1016/j.bulsci.2011.12.004. URL http://dx.doi.org/10.1016/j.bulsci.2011.12.004.

[65] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, pages 508–539, 1996.

[66] D. Dubhashi and O. Häggström. A note on conditioning and stochastic domination for order statistics. *J. Appl. Probab.*, 45(2):575–579, 2008. ISSN 0021-9002. doi: 10.1239/jap/1214950369. URL http://dx.doi.org/10.1239/jap/1214950369.

[67] C. F. Dunkl and Y. Xu. *Orthogonal polynomials of several variables*, volume 81. Cambridge University Press, 2001.

[68] F. Durante and P. Jaworski. Absolutely continuous copulas with given diagonal sections. *Communications in Statistics - Theory and Methods*, 37(18):2924–2942, 2008. doi: 10.1080/03610920802050927. URL http://www.tandfonline.com/doi/abs/10.1080/03610920802050927.

[69] F. Durante, R. Mesiar, and C. Sempi. On a family of copulas constructed from the diagonal section. *Soft Computing*, 10:490–494, 2006. ISSN 1432-7643. doi: 10.1007/s00500-005-0523-7. URL http://dx.doi.org/10.1007/s00500-005-0523-7.

[70] F. Durante, A. Kolesárová, R. Mesiar, and C. Sempi. Copulas with given diagonal sections: novel constructions and applications. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems*, 15(4):397–410, 2007. ISSN 0218-4885. doi: 10.1142/S0218488507004753. URL http://dx.doi.org/10.1142/S0218488507004753.

[71] F. Durante, J. F. Sánchez, and W. Trutschnig. Multivariate copulas with hairpin support. *Journal of Multivariate Analysis*, 130:323 – 334, 2014. ISSN 0047-259X. doi: http://dx.doi.org/10.1016/j.jmva.2014.06.009. URL http://www.sciencedirect.com/science/article/pii/S0047259X14001390.

[72] F. Durante, J. Fernández-Sánchez, and R. Pappadà. Copulas, diagonals, and tail dependence. *Fuzzy Sets and Systems*, 264:22–41, 2015.

[73] F. Durante, J. Fernández-Sánchez, J. J. Quesada-Molina, and M. Úbeda Flores. Diagonal plane sections of trivariate copulas. *Information Sciences*, 333:81 – 87, 2016. ISSN 0020-0255. doi: http://dx.doi.org/10.1016/j.ins.2015.11.024. URL http://www.sciencedirect.com/science/article/pii/S0020025515008348.

[74] S. Y. Efroimovich. Nonparametric estimation of a density of unknown smoothness. *Theory of Probability & Its Applications*, 30(3):557–568, 1986.

[75] V. A. Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability and Its Applications*, 14(1):153–158, 1969.

[76] A. Erdely and J. M. González-Barrios. On the construction of families of absolutely continuous copulas with given restrictions. *Comm. Statist. Theory Methods*, 35(4-6):649–659, 2006. ISSN 0361-0926. doi: 10.1080/03610920500498758. URL http://dx.doi.org/10.1080/03610920500498758.

[77] A. Erdely, J. M. González-Barrios, and M. M. Hernández-Cedillo. Frank's condition for multivariate archimedean copulas. *Fuzzy Sets and Systems*, 240:131 – 136, 2014. ISSN 0165-0114. doi: http://dx.doi.org/10.1016/j.fss.2013.05.017. URL http://www.sciencedirect.com/science/article/pii/S0165011413002595. Theme: Aggregation Operators.

[78] G. Fredricks and R. Nelsen. Copulas Constructed from Diagonal Sections. In V. Beneš and J. Štěpán, editors, *Distributions with given Marginals and Moment Problems*, pages 129–136. Springer Netherlands, 1997. ISBN 978-94-010-6329-6. doi: 10.1007/978-94-011-5532-8_16. URL http://dx.doi.org/10.1007/978-94-011-5532-8_16.

[79] D. Freedman and P. Diaconis. On the histogram as a density estimator: L_2 theory. *Probability Theory and Related Fields*, 57(4):453–476, 1981.

[80] D. Freedman and P. Diaconis. On the maximum deviation between the histogram and the underlying density. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 58 (2):139–167, 1981.

[81] T. Gasser, H. Muller, and V. Mammitzsch. Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 238–252, 1985.

[82] J. W. Gibbs, H. A. Bumstead, W. R. Longley, et al. *The collected works of J. Willard Gibbs*, volume 1. Longmans, Green and Company, 1928.

[83] E. Giné and R. Nickl. An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probability Theory and Related Fields*, 143(3):569–596, 2008. ISSN 1432-2064. doi: 10.1007/s00440-008-0137-y. URL http://dx.doi.org/10.1007/s00440-008-0137-y.

[84] A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608–1632, 2011.

[85] A. Goldenshluger and O. Lepski. On adaptive minimax density estimation on $\mathbb{R}^d$. *Probability Theory and Related Fields*, 159(3-4):479–543, 2014.

[86] G. K. Golubev. Nonparametric estimation of smooth densities of a distribution in L_2. *Problemy Peredachi Informatsii*, 28(1):52–62, 1992. ISSN 0555-2923.

[87] I. J. Good. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics*, pages 911–934, 1963.

[88] U. Grenander and G. Szegö. *Toeplitz forms and their applications*, volume 321. Univ of California Press, 1958.

[89] O. Grothe, J. Schnieders, and J. Segers. Measuring association and dependence between random vectors. *Journal of Multivariate Analysis*, 123:96–110, 2014.

[90] E. Guerre, I. Perrigne, and Q. Vuong. Optimal nonparametric estimation of first-price auctions. *Econometrica*, 68(3):525–574, 2000. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2999600.

[91] P. Hall. Estimating a density on the positive half line by the method of orthogonal series. *Annals of the Institute of Statistical Mathematics*, 32(1):351–362, 1980.

[92] P. Hall. On trigonometric series estimates of densities. *Ann. Statist.*, 9(3):683–685, 05 1981. doi: 10.1214/aos/1176345474. URL http://dx.doi.org/10.1214/aos/1176345474.

[93] P. Hall. Comparison of two orthogonal series methods of estimating a density and its derivatives on an interval. *Journal of Multivariate Analysis*, 12(3):432 – 449, 1982. ISSN 0047-259X. doi: http://dx.doi.org/10.1016/0047-259X(82)90076-8. URL http://www.sciencedirect.com/science/article/pii/0047259X82900768.

[94] P. Hall. On Kullback-Leibler loss and density estimation. *The Annals of Statistics*, pages 1491–1519, 1987.

[95] P. Hall and J. Marron. On the amount of noise inherent in bandwidth selection for a kernel density estimator. *The Annals of Statistics*, pages 163–181, 1987.

[96] P. Hall, G. Kerkyacharian, and D. Picard. Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.*, 26(3):922–942, 06 1998. doi: 10.1214/aos/1024691082. URL http://dx.doi.org/10.1214/aos/1024691082.

[97] Z. Hao and V. Singh. Entropy-copula method for single-site monthly streamflow simulation. *Water Resources Research*, 48(6), 2012.

[98] T. Herbst. An application of randomly truncated data models in reserving IBNR claims. *Insurance: Mathematics and Economics*, 25(2):123–131, 1999.

[99] T. Hu and H. Chen. Dependence properties of order statistics. *Journal of Statistical Planning and Inference*, 138(7):2214–2222, 2008.

[100] A. J. Izenman. Review papers: recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, 86(413):205–224, 1991.

[101] K. D. Jarman and P. D. Whitney. Integrating correlated bayesian networks using maximum entropy. *Applied Mathematical Sciences*, 5(48):2361–2371, 2011.

[102] P. Jaworski. On copulas and their diagonals. *Information Sciences*, 179(17):2863 – 2871, 2009. ISSN 0020-0255. doi: 10.1016/j.ins.2008.09.006. URL http://www.sciencedirect.com/science/article/pii/S0020025508003836.

[103] P. Jaworski and T. Rychlik. On distributions of order statistics for absolutely continuous copulas with applications to reliability. *Kybernetika*, 44(6):757–776, 2008.

[104] E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[105] H. Joe. *Multivariate models and dependence concepts*, volume 73 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1997. ISBN 0-412-07331-5.

[106] P. Joly, D. Commenges, and L. Letenneur. A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence of dementia. *Biometrics*, pages 185–194, 1998.

[107] A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric regression. *Ann. Statist.*, 28(3):681–712, 05 2000. doi: 10.1214/aos/1015951994. URL http://dx.doi.org/10.1214/aos/1015951994.

[108] A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *Ann. Statist.*, 36(5):2183–2206, 10 2008. doi: 10.1214/07-AOS546. URL http://dx.doi.org/10.1214/07-AOS546.

[109] G. Kerkyacharian and D. Picard. Density estimation in besov spaces. *Statistics and Probability Letters*, 13(1):15 – 24, 1992. ISSN 0167-7152. doi: http://dx.doi.org/10.1016/0167-7152(92)90231-S. URL http://www.sciencedirect.com/science/article/pii/016771529290231S.

[110] G. Kerkyacharian and D. Picard. Density estimation by kernel and wavelets methods: Optimality of besov spaces. *Statistics and Probability Letters*, 18(4):327 – 336, 1993. ISSN 0167-7152. doi: http://dx.doi.org/10.1016/0167-7152(93)90024-D. URL http://www.sciencedirect.com/science/article/pii/016771529390024D.

[111] G. Kerkyacharian, D. Picard, and K. Tribouley. L_p adaptive density estimation. *Bernoulli*, pages 229–247, 1996.

[112] G. Kerkyacharian, O. Lepski, and D. Picard. Nonlinear estimation in anisotropic multi-index denoising. *Probability theory and related fields*, 121(2):137–170, 2001.

[113] J. M. Keynes. *A treatise on probability*. Courier Corporation, 2013.

[114] S. Kim and H. David. On the dependence structure of order statistics and concomitants of order statistics. *Journal of statistical planning and inference*, 24(3):363–368, 1990.

[115] J. P. Klein and M. L. Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science and Business Media, 2005.

[116] J.-Y. Koo and W.-C. Kim. Wavelet density estimation by approximation of log-densities. *Statistics & probability letters*, 26(3):271–278, 1996.

[117] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[118] S. W. Lagakos, L. Barraj, and V. De Gruttola. Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika*, 75(3):515–523, 1988.

[119] R. Lebrun and A. Dutfoy. Copulas for order statistics with prescribed margins. *Journal of Multivariate Analysis*, 128:120–133, 2014.

[120] G. Lecué. Lower bounds and aggregation in density estimation. *The Journal of Machine Learning Research*, 7:971–981, 2006.

[121] G. Lecué and S. Mendelson. Aggregation via empirical risk minimization. *Probability Theory and Related Fields*, 145(3-4):591–613, 2009.

[122] O. Lepski et al. Multivariate density estimation under sup-norm loss: oracle approach, adaptation and independence structure. *The Annals of Statistics*, 41(2):1005–1034, 2013.

[123] O. V. Lepski. A problem of adaptive estimation in gaussian white noise. *Teoriya Veroyatnostei i ee Primeneniya*, 35(3):459–470, 1990.

[124] O. V. Lepskii. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1992. doi: 10.1137/1136085. URL http://dx.doi.org/10.1137/1136085.

[125] X. Luo and W.-Y. Tsai. Nonparametric estimation of bivariate distribution under right truncation with application to panic disorder. *Journal of Statistical Planning and Inference*, 139(4):1559–1568, 2009.

[126] D. Lynden-Bell. A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Notices of the Royal Astronomical Society*, 155(1):95–118, 1971.

[127] J. S. Marron and D. Nolan. Canonical kernels for density estimation. *Statistics and Probability Letters*, 7(3):195–199, 1988.

[128] A. Meeuwissen and T. Bedford. Minimally informative distributions with given rank correlation for use in uncertainty analysis. *Journal of Statistical Computation and Simulation*, 57(1-4):143–174, 1997. doi: 10.1080/00949659708811806. URL http://www.tandfonline.com/doi/abs/10.1080/00949659708811806.

[129] C. C. Moore. The degree of randomness in a stationary time series. *Ann. Math. Statist.*, 34:1253–1258, 1963.

[130] J. Navarro and N. Balakrishnan. Study of some measures of dependence between order statistics and systems. *Journal of Multivariate Analysis*, 101(1):52–67, 2010.

[131] J. Navarro and F. Spizzichino. On the relationships between copulas of order statistics and marginal distributions. *Statist. Probab. Lett.*, 80(5-6):473–479, 2010. ISSN 0167-7152. doi: 10.1016/j.spl.2009.11.025. URL http://dx.doi.org/10.1016/j.spl.2009.11.025.

[132] R. B. Nelsen. *An introduction to copulas*. Springer Series in Statistics. Springer, New York, second edition, 2006. ISBN 978-0387-28659-4; 0-387-28659-4.

[133] R. B. Nelsen, J. J. Q. Molina, J. A. R. Lallena, and M. Úbeda Flores. Best-possible bounds on sets of bivariate distribution functions. *Journal of Multivariate Analysis*, 90(2): 348 – 358, 2004. ISSN 0047-259X. doi: 10.1016/j.jmva.2003.09.002. URL http://www.sciencedirect.com/science/article/pii/S0047259X0300157X.

[134] R. B. Nelsen, J. J. Quesada-Molina, J. A. Rodríguez-Lallena, and M. Úbeda Flores. On the construction of copulas and quasi-copulas with given diagonal sections. *Insurance: Mathematics and Economics*, 42(2):473 – 483, 2008. ISSN 0167-6687. doi: 10.1016/j.insmatheco.2006.11.011. URL http://www.sciencedirect.com/science/article/pii/S0167668706001867.

[135] A. Nemirovski. Topics in non-parametric statistics. *Ecole d'Eté de Probabilités de Saint-Flour*, 28:85, 2000.

[136] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962. ISSN 00034851. URL http://www.jstor.org/stable/2237880.

[137] A. Pasanisi. *Uncertainty analysis and decision-aid: methodological, technical and managerial contributions to engineering and R&D studies*. PhD thesis, Université de Technologie de Compiègne, 2014.

[138] A. Pasanisi and A. Dutfoy. An industrial viewpoint on uncertainty quantification in simulation: Stakes, methods, tools, examples. In *Uncertainty Quantification in Scientific Computing*, pages 27–45. Springer, 2012.

[139] E. Pasha and S. Mansoury. Determination of maximum entropy multivariate probability distribution under some constraints. *Applied Mathematical Sciences*, 2(57):2843–2849, 2008.

[140] M. Petelet, B. Iooss, O. Asserin, and A. Loredo. Latin hypercube sampling with inequality constraints. *AStA Advances in Statistical Analysis*, 94(4):325–339, 2010.

[141] J. Piantadosi, P. Howlett, and J. Borwein. Copulas with maximum entropy. *Optimization Letters*, 6:99–125, 2012. ISSN 1862-4472. doi: 10.1007/s11590-010-0254-2. URL http://dx.doi.org/10.1007/s11590-010-0254-2.

[142] J. Piantadosi, P. Howlett, J. Borwein, and J. Henstridge. Maximum entropy methods for generating simulated rainfall. *Numerical Algebra, Control and Optimization*, 2(2):233–256, 2012. ISSN 2155-3289. doi: 10.3934/naco.2012.2.233. URL http://aimsciences.org/journals/displayArticlesnew.jsp?paperID=7385.

[143] H. Poincaré. *Calcul des probabilités*. Les Grands Classiques Gauthier-Villars. [Gauthier-Villars Great Classics]. Éditions Jacques Gabay, Sceaux, 1987. ISBN 2-87647-001-2. Reprint of the second (1912) edition.

[144] A.-L. Popelin, A. Balmont, J.-C. Clement, and J. Angles. Sensitivity analysis of a welding thermomechanical simulation model. In *Proceedings of the 12th Annual Conference of the European Network for Business and Industrial Statistics, Ljubjana, Slovenia, 2012*, 2012.

[145] A.-L. Popelin, A. Balmont, J.-C. Clement, and J. Angles. Sensitivity analysis of a numerical welding simulation model. In *Proceedings of the 21ème Congrès Français de Mécanique, Bordeaux, France, 2013*. AFM, 2013. URL http://hdl.handle.net/2042/52865.

[146] D.-B. Pougaza, A. Mohammad-Djafari, and J.-F. Bercher. Link between copula and tomography. *Pattern Recognition Letters*, 31(14):2258–2264, 2010.

[147] J. J. Quesada-Molina, S. Saminger-Platz, and C. Sempi. Quasi-copulas with a given subdiagonal section. *Nonlinear Analysis: Theory, Methods & Applications*, 69(12):4654 – 4673, 2008. ISSN 0362-546X. doi: 10.1016/j.na.2007.11.021. URL http://www.sciencedirect.com/science/article/pii/S0362546X07007754.

[148] E. Remy, A.-L. Popelin, and A. Feng. Modelling dependence using copulas - an implementation in the field of structural reliability. 2012. Paper presented at PSAM 11 and ESREL 2012 Conference on Probabilistic Safety Assessment.

[149] P. Rigollet. Kullback–Leibler aggregation and misspecified generalized linear models. *The Annals of Statistics*, 40(2):639–665, 2012.

[150] P. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, 16(3):260–280, 2007.

[151] M. Rosenblatt. Remarks on a multivariate transformation. *Ann. Math. Statist.*, 23(3): 470–472, 09 1952. doi: 10.1214/aoms/1177729394. URL http://dx.doi.org/10.1214/aoms/1177729394.

[152] M. Rosenblatt et al. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

[153] L. Rüschendorf and W. Thomsen. Note on the Schrödinger equation and I-projections. *Statist. Probab. Lett.*, 17(5):369–375, 1993. ISSN 0167-7152. doi: 10.1016/0167-7152(93)90257-J. URL http://dx.doi.org/10.1016/0167-7152(93)90257-J.

[154] T. Rychlik. Distributions and expectations of order statistics for possibly dependent random variables. *Journal of Multivariate Analysis*, 48(1):31 – 42, 1994. ISSN 0047-259X. doi: http://dx.doi.org/10.1016/0047-259X(94)80003-E. URL http://www.sciencedirect.com/science/article/pii/0047259X9480003E.

[155] A. Samarov and A. Tsybakov. Aggregation of density estimators and dimension reduction. *Advances in statistical modeling and inference. Essays in honor of Kjell A. Doksum*, pages 233–251, 2007.

[156] V. Schmitz. Revealing the dependence structure between $X_{(1)}$ and $X_{(n)}$. *Journal of statistical planning and inference*, 123(1):41–47, 2004.

[157] P. J. Schönbucher. *Credit Derivatives pricing models: models, pricing, and implementation*. John Wiley & Sons, 2003.

[158] S. C. Schwartz. Estimation of probability density by an orthogonal series. *Ann. Math. Statist.*, 38(4):1261–1265, 08 1967. doi: 10.1214/aoms/1177698795. URL http://dx.doi.org/10.1214/aoms/1177698795.

[159] D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

[160] D. W. Scott and G. R. Terrell. Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, 82(400):1131–1146, 1987.

[161] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[162] M. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

[163] E. Strzalkowska-Kominiak and W. Stute. On the probability of holes in truncated samples. *Journal of Statistical Planning and Inference*, 140(6):1519 – 1528, 2010. ISSN 0378-3758. doi: http://dx.doi.org/10.1016/j.jspi.2009.12.017. URL http://www.sciencedirect.com/science/article/pii/S0378375809003978.

[164] E. A. Sungur and Y. Yang. Diagonal copulas of Archimedean class. *Comm. Statist. Theory Methods*, 25(7):1659–1676, 1996. ISSN 0361-0926. doi: 10.1080/03610929608831791. URL http://dx.doi.org/10.1080/03610929608831791.

[165] G. R. Terrell and D. W. Scott. Variable kernel density estimation. *Ann. Statist.*, 20(3): 1236–1265, 09 1992. doi: 10.1214/aos/1176348768. URL http://dx.doi.org/10.1214/aos/1176348768.

[166] R. C. Tolman. *The principles of statistical mechanics*. Oxford University Press, 1938.

[167] A. Tsybakov. Aggregation and high-dimensional statistics. *Lecture notes for the course given at the École d'été de Probabilités in Saint-Flour, URL: http://www.crest.fr/ckfinder/userfiles/files/Pageperso/ATsybakov/Lecture_notes_SFlour.pdf*, 2013.

[168] A. B. Tsybakov. Optimal rates of aggregation. In B. Schölkopf and M. K. Warmuth, editors, *Learning Theory and Kernel Machines*, volume 2777 of *Lecture Notes in Computer Science*, pages 303–313. Springer Berlin Heidelberg, 2003. ISBN 978-3-540-40720-1. doi: 10.1007/978-3-540-45167-9_23. URL http://dx.doi.org/10.1007/978-3-540-45167-9_23.

[169] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

[170] B. W. Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B*, 38(3):290–295, 1976. ISSN 0035-9246.

[171] T. Wagner. Nonparametric estimates of probability densities. *IEEE Transactions on Information Theory*, 21(4):438–440, Jul 1975. ISSN 0018-9448. doi: 10.1109/TIT.1975.1055408.

[172] G. Wahba. Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation. *Ann. Statist.*, 3(1):15–29, 01 1975. doi: 10.1214/aos/1176342997. URL http://dx.doi.org/10.1214/aos/1176342997.

[173] G. G. Walter. Approximation of the delta function by wavelets. *Journal of Approximation Theory*, 71(3):329 – 343, 1992. ISSN 0021-9045. doi: http://dx.doi.org/10.1016/0021-9045(92)90123-6. URL http://www.sciencedirect.com/science/article/pii/0021904592901236.

[174] M. Wegkamp. Model selection in nonparametric regression. *Ann. Statist.*, 31(1):252–273, 02 2003. doi: 10.1214/aos/1046294464. URL http://dx.doi.org/10.1214/aos/1046294464.

[175] M. H. Wegkamp. Quasi-universal bandwidth selection for kernel density estimators. *Canadian Journal of Statistics*, 27(2):409–420, 1999.

[176] X. Wu. Calculation of maximum entropy densities with application to income distribution. *Journal of Econometrics*, 115(2):347–354, 2003.

[177] X. Wu. Exponential series estimator of multivariate densities. *Journal of Econometrics*, 156(2):354–366, 2010.

[178] Y. Yang. Combining different procedures for adaptive regression. *Journal of Multivariate Analysis*, 74(1):135–161, 2000.

[179] Y. Yang. Mixing strategies for density estimation. *The Annals of Statistics*, 28(1):75–87, 2000.

[180] Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10(1): 25–47, 02 2004. doi: 10.3150/bj/1077544602. URL http://dx.doi.org/10.3150/bj/1077544602.

[181] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999. ISSN 00905364. URL http://www.jstor.org/stable/2674082.

[182] T. Zhang et al. From ε-entropy to kl-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.

[183] N. Zhao and W. T. Lin. A copula entropy approach to correlation measurement at the country level. *Applied Mathematics and Computation*, 218(2):628 – 642, 2011. ISSN 0096-3003. doi: 10.1016/j.amc.2011.05.115. URL http://www.sciencedirect.com/science/article/pii/S0096300311007983.