Policy evaluation, high-dimension and machine learning

Jérémy L'Hour

université PARIS-SACLAY

ENSAE

IP PARIS

NNT : 2019SACLG008

Thèse de doctorat

Évaluation des politiques publiques, grande dimension, et machine learning

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

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

Thèse présentée et soutenue à Palaiseau, le 13 décembre 2019, par

JÉRÉMY L'HOUR

Composition du Jury :

Luc Behaghel Directeur de Recherche INRA, Paris School of Economics	Président et Rapporteur
Christoph Rothe Professeur, University of Mannheim	Rapporteur
Alberto Abadie Professeur, Massachusetts Institute of Technology (MIT)	Examinateur
Victor-Emmanuel Brunel Assistant Professeur, ENSAE Paris (CREST)	Examinateur
Xavier D'Haultfœuille Professeur, ENSAE Paris (CREST)	Directeur de thèse

Remerciements

Je remercie d'abord mon directeur de thèse, Xavier D'Haultfœuille. Je lui suis très reconnaissant pour la liberté qu'il m'a laissé de définir mes thèmes de recherche, sa disponibilité, ses encouragements bienvenus lorsque je désespérais d'une preuve qui n'aboutissait pas, et les nombreuses discussions que nous avons eues. Grâce à lui, je crois que je commence à comprendre l'économétrie. Cette thèse se termine, mais pas notre collaboration, je l'espère.

Alberto Abadie m'a gentiment accueilli à Harvard, puis au MIT par deux reprises. Je le remercie chaleureusement pour le temps qu'il m'a consacré, à la fois au bureau et en dehors, sa bienveillance et l'ambition qu'il a pour notre projet commun.

Je remercie Luc Behaghel, Christoph Rothe et Victor-Emmanuel Brunel d'avoir accepté de participer à mon jury de thèse et d'avoir de ce fait produit des commentaires pertinents pour faire progresser ma recherche.

Cette thèse est essentiellement le fruit de collaborations. Je remercie donc mes (autres) co-auteurs: Marianne Bléhaut, Bruno Crépon, Esther Duflo, Elia Pérennès et Sacha Tsybakov. Je tiens à remercier tous les collègues du CREST et membres de la communauté statistique et économétrique qui ont pris du temps pour discuter, échanger des idées, partager leurs savoirs ou me relire. Cette thèse serait d'une qualité moindre sans ces rencontres. Je pense plus particulièrement à Pierre Alquier, Gauthier Appert, Espen Bernton, Victor-Emmanuel Brunel, Alexander Buchholz, Léna Carel, Arthur Cazaubiel, Badr-Eddine Cherief-Adbellatif, Geoffrey Chinot, Nicolas Chopin, Vincent Cottet, Morgane Cure, Marco Cuturi, Arnak Dalalyan, Laurent Davezies, Avi Feller, Christophe Gaillac, Lucas Girard, Malka Guillot, Yannick Guyonvarch, Guillaume Lecué, Clémence Lenoir, Boris Muzellec, Simo Ndaoud, Julie Pernaudet, Audrey Rain, Lionel Riou-Durand, Anna Simoni, Jann Spiess, Jérôme Trinh, Ao Wang et Meriem Zaiem. Je remercie Josh Angrist, Stéphane Bonhomme, Richard Blundell, Matias Cattaneo, Clément de Chaisemartin, Max Farrell, Laurent Gobillon, Stefan Hoderlein, Hyunseung Kang, Philipp Ketz, Guido Imbens, Arthur Lewbel, David Margolis, Jamie Robins et Kaspar Wuthrich – et j'en oublie certainement – qui ont accepté de me rencontrer au cours de ces quatre dernières années pour critiquer mon travail.

Je remercie Francis Kramarz ainsi que tout le personnel du CREST, et tiens à dire combien je suis fier d'avoir effectué ma thèse au CREST, dont le positionnement à l'intersection des statistiques mathématiques et de l'économie m'a été bénéfique. J'espère que cette thèse en illustre en partie l'essence. Je remercie également mes camarades

doctorants des laboratoires de statistique et d'économie, ainsi que l'ENSAE et tous les collègues de la direction des études.

Enfin, je remercie du fond du cœur les professeurs que j'ai rencontrés durant mon parcours scolaire et qui ont cru en moi. Je suis éternellement reconnaissant à Isabelle Morlais de m'avoir soutenu. Je remercie tous mes amis, ainsi que ma sœur Isabelle, mon frère Kévin et mon père Denis. Ils ont tenu un rôle silencieux mais essentiel dans l'aboutissement de ce travail. Je dédie ce manuscrit à ma mère, Jocelyne, qui nous a quittés beaucoup trop tôt.

Contents

0 Résumé substantiel en Français **1**

1 Introduction **3**

1 High-Dimension, Variable Selection and Immunization 3

1.1 The Post-Selection Inference Problem 4

1.2 State of the Art 7

1.3 Contribution 9

2 Machine Learning in Empirical Economics 10

2.1 State of the Art 10

2.2 Contribution 11

2.3 Perspectives 12

3 Synthetic Control, High-Dimension and Selection of the Control Group . 14

3.1 State of the Art 14

3.2 Contribution 15

3.3 Perspectives 16

2 A Parametric Alternative to the Synthetic Control Method with Many Covariates **19**

1 Introduction 20

2 A Parametric Alternative to Synthetic Control 22

2.1 Covariate Balancing Weights and Double Robustness 22

2.2 Asymptotic Properties in Low-Dimension 24

3 High-Dimensional Covariates and Post-Selection Inference 25

3.1 Regularized Estimation 25

3.2 Immunized Estimation 27

3.3 Asymptotic Properties 28

4 Simulations . . . 31
5 Empirical Applications . . . 38
5.1 Job Training Program, LaLonde (1986) . . . 38
5.2 California Tobacco Control Program, Abadie et al. (2010) . . . 39
6 Conclusion . . . 42
7 Appendix A: Algorithm for Feasible Penalty Loadings . . . 43
8 Appendix B: Proofs . . . 43

3 A Penalized Synthetic Control Estimator for Disaggregated Data 63
1 Introduction . . . 64
2 Penalized Synthetic Control . . . 66
2.1 Synthetic Control for Disaggregated Data . . . 66
2.2 Penalized Synthetic Control . . . 69
2.3 Bias-Corrected Synthetic Control . . . 73
3 Large Sample Properties . . . 74
3.1 Bias . . . 74
3.2 Consistency . . . 74
3.3 Asymptotic Normality . . . 75
3.4 Asymptotic Behavior of $S(\lambda)$ . . . 76
4 Permutation Inference . . . 77
4.1 Inference on Aggregate Effects . . . 77
4.2 Inference Based on the Sum of Rank Statistics of Unit-Level Treatment Effects Estimates . . . 78
5 Penalty Choice . . . 79
5.1 Leave-One-Out Cross-Validation of Post-Intervention Outcomes for the Untreated . . . 80
5.2 Pre-Intervention Holdout Validation on the Outcomes of the Treated 80
6 Simulations . . . 81
7 Empirical Applications . . . 83
7.1 The Value of Connections in Turbulent Times, Acemoglu et al. (2016) 83
7.2 The Impact of Election Day Registration on Voter Turnout, Xu (2017) . . . 88
8 Conclusion . . . 91
9 Appendix: Proofs . . . 95

4 Using Generic Machine Learning to Analyze Treatment Heterogeneity: An Application to Provision of Job Counseling **107**

1 Introduction . . . 108
2 Machine Learning in Empirical Economics . . . 109
3 Data and Experimental Design . . . 110
3.1 Design of the Experiment . . . 111
3.2 Data . . . 111
4 Empirical Strategy . . . 112
4.1 An Economic Model of Treatment Allocation . . . 112
4.2 Methodological Aspects . . . 114
5 Results . . . 115
5.1 Detection of Heterogeneity . . . 116
5.2 Dimension of Heterogeneity (CLAN) . . . 118
5.3 Selection into the Treatment . . . 121
6 Conclusion . . . 121
7 Appendix: Descriptive Statistics . . . 127
8 Appendix: Adaptation of Th. 2.1 in Chernozhukov et al. (2018b) . . . 127

Bibliography **129**

Chapter 0

Résumé substantiel en Français

Cette thèse regroupe trois travaux d'économétrie reliés par l'application de l'apprentissage automatique et de la statistique en grande dimension à l'évaluation de politiques publiques et l'inférence causale.

La première partie propose une alternative paramétrique au contrôle synthétique (Abadie and Gardeazabal, 2003; Abadie et al., 2010) dans un contexte de grande dimension du vecteur des caractéristiques à apparier. Il prend la forme d'un estimateur reposant sur une première étape de type Lasso, et on montre qu'il est doublement robuste, asymptotiquement Normal et "immunisé" contre les erreurs de première étape. On montre qu'il permet de palier les limitations du contrôle synthétique, en particulier dans le cadre de données micro-économiques. En particulier, l'estimateur proposé donne une solution unique qui satisfait aux mêmes contraintes que le contrôle synthétique (non-négativité et somme à un). Il permet également d'opérer une sélection de variable sur le même mode que le Lasso, ce qui donne une alternative à la méthode originale du contrôle synthétique offrant une procédure qui optimise la matrice V de pondération des variables via la norme $\|.\|_V$ afin d'améliorer l'adéquation à la tendance pré-traitement. Finalement, on montre que l'immunisation aboutit à une procédure de correction du biais lorsque le calage sur une variable n'est pas parfait et que cette variable est pertinente pour prédire le résultat sans le traitement.

La seconde partie étudie une version pénalisée du contrôle synthétique pour des données de nature micro-économique. La pénalisation permet d'obtenir une unité synthétique qui réalise un arbitrage entre, d'une part, reproduire fidèlement l'unité traitée durant la période pré-traitement et, d'autre part, n'utiliser que des unités non-traitées suffisamment semblables à celles-ci. Lorsque la pénalisation est suffisamment forte, l'estimateur du contrôle synthétique pénalisé coïncide avec l'estimateur de *matching* au plus proche voisin. Nous établissons les propriétés géométriques de la solution ainsi que les propriétés asymptotiques de l'estimateur. Enfin, nous proposons deux procédures de type "validation croisée" qui permettent de choisir la pénalisation et discutons des procédures d'inférence par permutation dans ce contexte.

Ces deux chapitres, de nature théorique, sont complétés par des simulations ainsi que des replications d'articles empiriques. Le premier chapitre utilise des données d'un pro-

gramme américain d'emplois aidés ciblant les personnes éloignées du marché du travail, et revisite également la "Proposition 99", un programme de lutte contre le tabac mis en place en Californie à la fin des années 1980, en tentant de quantifier son impact sur la consommation individuelle de tabac. Le second chapitre utilise des données boursières pour quantifier l'impact de la nomination de Tim Geithner à la tête du Trésor américain au cœur de la crise de 2008 sur la valorisation des firmes auxquelles il était connecté, et des données sur de participation aux élections présidentielles américaines au XXème siècle pour quantifier l'impact d'une loi permettant de s'inscire sur les listes électorales le jour du vote.

La dernière partie, à dominante empirique, porte sur l'application du *Generic Machine Learning* (Chernozhukov et al., 2018b) afin d'étudier l'hétérogénéité des effets d'une expérience aléatoire visant à comparer la fourniture publique et privée d'aide à la recherche d'emploi. Elle utilise les données d'une expérience française conduite par Pôle Emploi en 2007-2008. D'un point de vue méthodologique, ce projet discute l'extension du *Generic Machine Learning* à des expériences avec *compliance* imparfaite.

Chapter 1

Introduction

This introductory chapter presents the common themes that glue together the three core parts of this PhD dissertation. Within each section, I describe the main intuitions and the state of the art, I summarize the contribution of this thesis and outline a few perspectives on the topic.

Section 1 introduces the post-selection inference problem and its solution in the form of the double selection method that paved the way to the integration of Machine Learning (ML) to the empirical researcher's toolkit. Building on this first part, Section 2 discusses the value-added of ML for econometricians, in particular in the context of Randomized Controlled Trials (RCT). Finally, Section 3 discusses the synthetic control method, a popular tool in policy evaluation, which has a close connection with the high-dimensional literature and shares the same spirit of "letting the data speak" as ML algorithms.

1. High-Dimension, Variable Selection and Immunization

Model selection and parsimony among explanatory variables are traditional scientific problems that have a particular echo in Econometrics. Leamer (1983) was one of the first to raise the issue, famously quipping: *"there are two things you are better off not watching in the making: sausages and econometric estimates"*, thereby questioning the arbitrariness characterizing model choices. As high-dimensional datasets have become increasingly available to statisticians in various fields, model selection has garnered growing attention over the past two decades. But even with a small dataset, high-dimensional problems can occur, for example when doing series estimation of a non-parametric model. In practice, applied researchers select variables by trial and error, guided by their intuition and report results based on the assumption that the selected model is the true. These results are often backed by further sensitivity analysis and robustness checks. However, the variable selection step of empirical work is rarely fully acknowledged although it is not innocuous.

The so-called *post-selection inference problem* is the stepping stone to understand the recent developments in the Econometrics of high-dimensional models, all the way to the integration of ML algorithms to the applied researcher's toolkit. The intuition is simple

and carries over to more complicated cases where the first step does not necessarily involve selecting variables, but relies on estimation of an unknown quantity using a non-standard statistical tool such as a ML algorithm. The first section reviews the intuitions of Leeb and Pötscher (2005). The second section illustrates the solution developed, for example, in Belloni et al. (2014a). The third section discusses how this method is applied in the first chapter of this thesis to propose an alternative to the synthetic control method.

1.1. The Post-Selection Inference Problem

I begin by analyzing the two-step inference method, *i.e.* selecting the model first, then reporting results from that model as if it were the truth, in a small-dimensional context. The intuition easily extends to the high-dimensional case. This section is based on the work of Leeb and Pötscher (2005).

Assumption 1.1 (Possibly Sparse Gaussian Linear Model) *Consider the independent and identically distributed sequence of random variables* $(Y_i, X_i)_{i=1,\dots,n}$ *such that:*

$$Y_i = X_{i,1}\tau_0 + X_{i,2}\beta_0 + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, σ^2 *is known,* $X_i = (X_{i,1}, X_{i,2})$ *is a vector of dimension 2,* $\varepsilon_i \perp\!\!\!\perp X_i$, *and* $\mathbb{E}(X_i X_i')$ *is non-singular. I use the following shorthand notation for the OLS variance-covariance matrix elements:*

$$\begin{bmatrix} \sigma_\tau^2 & \sigma_{\tau,\beta} \\ \sigma_{\tau,\beta} & \sigma_\beta^2 \end{bmatrix} := \sigma^2 \left[\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right]^{-1}.$$

The most sparse true model is coded by M_0, *a random variable taking value* R *("restricted") if* $\beta_0 = 0$ *and* U *("unrestricted") otherwise.*

The econometrician is interested in performing inference over the parameter τ_0 and wonders whether he should include $X_{i,2}$ in the regression. At the end, he reports the result from model $\hat{M}$ he has selected in a first step. In policy evaluation, $X_{i,1}$ is typically the treatment of interest and $X_{i,2}$ a control variable. I denote by $\hat{\tau}(U)$ and $\hat{\beta}(U)$ the OLS estimators in the unrestricted model (model U) and by $\hat{\tau}(R)$ and $\hat{\beta}(R) = 0$ the restricted OLS estimators (model R).

Everything in this section will be conditional on the covariates $(X_i)_{1 \leq i \leq n}$ but I leave that dependency hidden. In particular, conditional on the covariates, the unrestricted estimator is Normally distributed:

$$\sqrt{n} \begin{bmatrix} \hat{\beta}(U) - \beta_0 \\ \hat{\tau}(U) - \tau_0 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\beta^2 & \sigma_{\tau,\beta} \\ \sigma_{\tau,\beta} & \sigma_\tau^2 \end{bmatrix} \right).$$

The econometrician includes $X_{i,2}$ in the regression if its corresponding Student statistics is large enough:

Assumption 1.2 (Decision Rule)

$$\hat{M} = \begin{cases} U & \text{if } |\sqrt{n}\hat{\beta}(U)/\sigma_\beta| > c_n \\ R & \text{otherwise,} \end{cases}$$

with $c_n \to \infty$ *and* $c_n/\sqrt{n} \to 0$ *as* $n \to \infty$.

The AIC criterion corresponds to $c_n = \sqrt{2}$ and the BIC to $c_n = \sqrt{\log n}$. How does this selection method performs asymptotically?

Lemma 1.1 (Model Selection Consistency) *For* $M_0 \in \{U, R\}$,

$$\mathbb{P}_{M_0}\left(\hat{M} = M_0\right) \to 1,$$

as $n \to \infty$, *where* $\mathbb{P}_{M_0}$ *indicates the probability distribution of* $\hat{M}$ *under the true model* M_0.

All the proofs can be found in Gaillac and L'Hour (2019) or in the original paper of Leeb and Pötscher (2005). Since the probability of selecting the true model tends to one with the sample size, Lemma 1.1 might induce you to think that a consistent model selection procedure allows inference to be performed "as usual", *i.e.* that the model selection step can be overlooked. However, for any given sample size n, the probability of selecting the true model can be very small if β_0 is close to zero without exactly being zero. For example, assume that $\beta_0 = \delta\sigma_\beta c_n/\sqrt{n}$ with $|\delta| < 1$ then: $\sqrt{n}\beta_0/\sigma_\beta = \delta c_n$ and the probability in the proof of Lemma 1.1 is equal to $1 - \Phi(c_n(1+\delta)) + \Phi((\delta-1)c_n)$, and tends to zero although the true model is U because $\beta_0 \neq 0$! This quick analysis tells us that the model selection procedure is blind to small deviations from the restricted model ($\beta_0 = 0$) that are of the order of $c_n/\sqrt{n}$. Statisticians say that in that case, the model selection procedure is not *uniformly consistent* with respect to β_0. For the applied researcher, it means that the classical inference procedure, *i.e.* the procedure that assumes that the selected model is the true, or that is conditional on the selected model being the true, and uses the asymptotic normality to perform tests may require very large sample sizes to be accurate. Furthermore, this required sample size depends on the unknown parameter β_0.

More interestingly, Leeb and Pötscher (2005) analyze the distribution of the post-selection estimator $\tilde{\tau}$ defined by

$$\tilde{\tau} := \hat{\tau}(\hat{M}) = \hat{\tau}(R)\mathbf{1}_{\hat{M}=R} + \hat{\tau}(U)\mathbf{1}_{\hat{M}=U}.$$

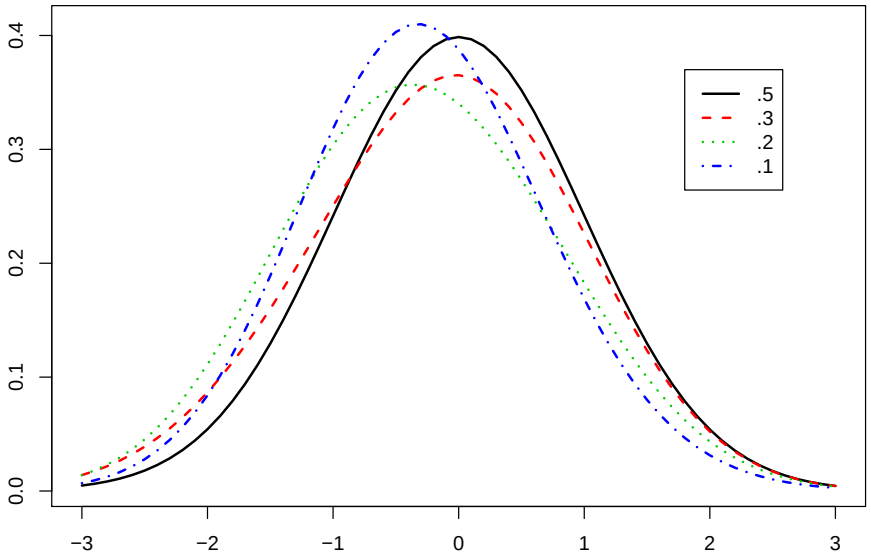
Bearing in mind the caveat issued in the previous paragraph, is a consistent model selection procedure sufficient to waive concerns over the post-selection approach? Indeed, using Lemma 1.1, it is tempting to think that, $\tilde{\tau}$ will be asymptotically distributed as a Gaussian and that standard asymptotic inference can be used to approximate the finite-sample behavior of the estimator. However, let us show that its finite-sample distribution can be very different from a standard Gaussian distribution.

Lemma 1.2 (Density of the Post-Selection estimator, from Leeb, 2006) *The finite-sample (conditional on $(X_i)_{i=1,\dots,n}$) density of $\sqrt{n}(\tilde{\tau}-\tau_0)$ is given by:*

$$f_{\sqrt{n}(\tilde{\tau}-\tau_0)}(x) = \Delta\left(\sqrt{n}\frac{\beta_0}{\sigma_\beta}, c_n\right)\frac{1}{\sigma_\tau\sqrt{1-\rho^2}}\varphi\left(\frac{x}{\sigma_\tau\sqrt{1-\rho^2}} + \frac{\rho}{\sqrt{1-\rho^2}}\frac{\sqrt{n}\beta_0}{\sigma_\beta}\right)$$
$$+\left[1-\Delta\left(\frac{\sqrt{n}\beta_0/\sigma_\beta+\rho x/\sigma_\tau}{\sqrt{1-\rho^2}}, \frac{c_n}{\sqrt{1-\rho^2}}\right)\right]\frac{1}{\sigma_\tau}\varphi\left(\frac{x}{\sigma_\tau}\right),$$

where $\rho = \sigma_{\tau,\beta}/\sigma_\tau\sigma_\beta$, $\Delta(a,b) := \Phi(a+b)-\Phi(a-b)$ and φ and Φ are the density and cdf of $\mathcal{N}(0,1)$, respectively.

Figure 1.1: Finite-sample density of $\sqrt{n}(\tilde{\tau}-\tau_0)$, $\rho = .4$



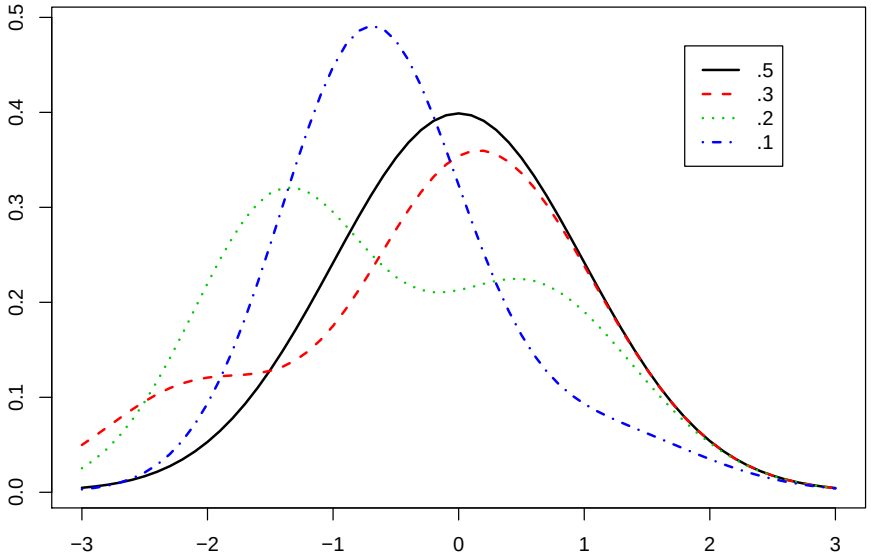
Note: Density of the post-selection estimator $\tilde{\tau}$ for different values of β_0/σ_β, see legend. Other parameters are: $c_n = \sqrt{\log n}$, $n = 100$, $\sigma_\tau = 1$ and $\rho = .4$. See Lemma 1.2 for the mathematical formula.

Notice that the bias corresponds to the usual omitted-variable bias since

$$-\beta_0\rho\sigma_\tau/\sigma_\beta \xrightarrow{p} \beta_0 \mathrm{Cov}(X_{i,1}, X_{i,2})/\mathrm{V}(X_{i,1}).$$

The fundamental problem is the omitted variable bias that the post-selection estimator cannot overcome unless $\beta_0 = 0$ or $\rho = 0$. Indeed, when $\rho = 0$, $\sqrt{n}(\tilde{\tau}-\tau_0) \sim \mathcal{N}(0,\sigma_\tau^2)$; while when $\beta_0 = 0$, $\sqrt{n}(\tilde{\tau}-\tau_0) \sim \mathcal{N}(0,\sigma_\tau^2/(1-\rho^2))$ (approximately), because $\Delta(0,c_n) \geq 1-\exp(-c_n^2/2)$ - the probability of selecting the restricted model - is large. Figures 1.1 and 1.2 plot the finite-sample density of the post-selection estimator for several values of β_0/σ_β

Figure 1.2: Finite-sample density of $\sqrt{n}(\tilde{\tau} - \tau_0)$, $\rho = .7$



Note: See Figure 1.1. $\rho = .7$.

in the cases $\rho = .4$ and $\rho = .7$, respectively. Figure 1.1 shows a mild albeit significant distortion from a standard Gaussian distribution. The post-selection estimator clearly exhibits a bias. As the correlation between the two covariates intensifies, the density of the post-selection estimator becomes highly non-Gaussian, even exhibiting two modes. Following this analysis, it is clear that inference based on standard Gaussian quantiles will in general give a picture very different from true distribution depicted in Figure 1.2.

1.2. State of the Art

Now, I suppose a high-dimensional set of control variable: in the model of the previous section, I assume that $p := \dim(X_{i,2})$ is large, possibly larger than the sample size n, *i.e.* β_0 is a high-dimensional nuisance parameter. Because high-dimensional nuisance parameters require the use of non-standard tools, they can disrupt the standard inference framework.

The message conveyed in the previous section was one of caution regarding the use of selection devices such as the Lasso in empirical Economics: inference, without taking into account the variable selection step, can be highly misleading. Even without explicit reference to variable selection, estimation of a high-dimensional nuisance parameter using ML algorithms will in general not lead to $\sqrt{n}$-consistent estimator and entails what is called a *regularization* bias in the second step (Belloni et al., 2014b).

Now, assume that β_0 is estimated using some ML algorithm, $\hat{\beta}$, unlikely to be $\sqrt{n}$-

consistent. Notice that the Normal equation implicitly used to define τ_0 in Assumption 1.1 is

$$E\left[\left(Y_i - X_{i,1}\tau_0 - X_{i,2}'\beta_0\right)X_{i,1}\right] = 0, \tag{1.1}$$

and the corresponding estimator takes the form

$$\begin{aligned}\hat{\tau} = \frac{\sum_{i=1}^n \left(Y_i - X_{i,2}'\hat{\beta}\right)X_{i,1}}{\sum_{i=1}^n X_{i,1}^2} &= \frac{\sum_{i=1}^n \left(Y_i - X_{i,2}'\beta_0\right)X_{i,1}}{\sum_{i=1}^n X_{i,1}^2} + \left(\beta_0 - \hat{\beta}\right)'\frac{\sum_{i=1}^n X_{i,2}X_{i,1}}{\sum_{i=1}^n X_{i,1}^2} \\ &= \tau_0 + \frac{\sum_{i=1}^n \varepsilon_i X_{i,1}}{\sum_{i=1}^n X_{i,1}^2} + \left(\beta_0 - \hat{\beta}\right)'\frac{\sum_{i=1}^n X_{i,2}X_{i,1}}{\sum_{i=1}^n X_{i,1}^2},\end{aligned}$$

for some ML estimator $\hat{\beta}$. The regularization bias comes from the fact that $\hat{\tau}$ is not first-order insensitive to errors in the estimation of β_0, *i.e.* the quantity $\sum_{i=1}^n X_{i,2}X_{i,1}/\sum_{i=1}^n X_{i,1}^2$ will, in general, converge in probably to a non-zero constant. This fact, combined with the fact that, in general, if $\hat{\beta}$ comes from a ML algorithm, we don't have $\sqrt{n}\left(\beta_0 - \hat{\beta}\right) = O_P(1)$ will mean that in general $\sqrt{n}\left(\hat{\tau} - \tau_0\right)$ will not be asymptotically Gaussian with mean zero. Of course, when β_0 is of small-dimension, this is not a problem because it suffices to replace $\hat{\beta}$ by an OLS estimator. But when p is very large, this may not be possible or even desirable.

The trick is to replace the moment equation (1.1) by another one, $E[\psi(Y_i, X_i, \tau_0, \eta_0)] = 0$, for some moment ψ and nuisance parameter η_0 such that $\partial_\eta E[\psi(Y_i, X_i, \tau_0, \eta_0)] = 0$. This last condition, which is an orthogonality condition, ensures that the estimating moment is first-order insensitive to deviations from the true value of the nuisance parameter. It helps "immunizing" the estimator of τ_0 from the first-step estimation using ML tools. In my example, ψ takes the form

$$E[\psi(Y_i, X_i, \tau, \eta)] = E\Bigg[\underbrace{\left(Y_i - X_{i,1}\tau - X_{i,2}'\beta\right)}_{\substack{\text{Residual from} \\ \text{Outcome Regression}}} \underbrace{\left(X_{i,1} - X_{i,2}'\delta\right)}_{\substack{\text{Residual from} \\ \text{Regression of } X_1 \text{ on } X_2}}\Bigg],$$

with $\eta = (\beta, \delta)$, and δ_0 such that $E\left[X_{i,2}\left(X_{i,1} - X_{i,2}'\delta_0\right)\right] = 0$. It is easy to check that $\partial_\beta E[\psi(Y_i, X_i, \tau_0, \eta_0)] = \partial_\delta E[\psi(Y_i, X_i, \tau_0, \eta_0)] = 0$. Notice that immunization also requires the estimation of another nuisance parameter, δ_0, which is of the same dimension as β_0. An estimation of τ_0 based on the previous equation will, under mild conditions, be asymptotically Gaussian – see Theorem 1 in Belloni et al. (2014b).

The previous equation has a Frish-Waugh-Lovell flavour. This idea has been developed and extended in many papers by Victor Chernozhukov and his co-authors, for example in Chernozhukov et al. (2015); Chernozhukov et al. (2015); Belloni et al. (2017); Chernozhukov et al. (2017, 2018a), and appears under the name of *double selection* when the problem comes from the use of the Lasso in a linear regression, *immunized* or *Neyman-orthogonal* estimators in a general framework or *double machine learning* when it is specifically applied to ML estimators. Typically, when using the Lasso, the procedure takes the following form:

1. Regress X_1 on X_2 using a Lasso, obtain $\hat{\delta}^L$. Define $\hat{S}_D := \left\{ j = 1, ..., p, \hat{\delta}_j^L \neq 0 \right\}$ the set of selected variables,
2. Regress Y on X_2 using a Lasso, obtain $\hat{\beta}^L$. Define $\hat{S}_Y := \left\{ j = 1, ..., p, \hat{\beta}_j^L \neq 0 \right\}$,
3. Regress Y on X_1 and the $\hat{s} = |\hat{S}_D \cup \hat{S}_Y|$ elements in X_2 which correspond to the indices $j \in \hat{S}_D \cup \hat{S}_Y$, using OLS.

And the resulting estimator in the last step will be asymptotically Gaussian. Although the post-selection inference problem is now well-understood and the tools to resolve it have been well developed, I believe that recalling the intuitions are important because they constitute a key step for the integration of ML methods in the empirical researcher's toolkit.

1.3. Contribution

In Chapter 2 of this thesis, we propose an alternative to the synthetic control method by suggesting a parametric form for the weights. The synthetic control method developed by Abadie et al. (2010) is a an econometric tool to evaluate causal effects in presence of a few treated units – see Section 3 below for a detailed presentation. While initially aimed at evaluating the impact of large-scale macroeconomic changes with very few available control units, it has increasingly been used in place of more well-known microeconometric tools in a broad range of applications, but its properties in this context are not completely known. In particular, when there are many untreated units, the minimization program defining the synthetic control weights becomes high-dimensional and an infinite number of solutions may exist.

Our parametric alternative is developed both in the usual asymptotic framework and in the high-dimensional one. The use of a parametric form for the weights allows to change the dimensionality of the problem from the number of untreated units to the number of covariates to balance between the treated and the control group, thereby building on the more standard high-dimensional statistics literature. In the low-dimensional context, *i.e.* when the number of covariates is fixed and much smaller than the sample size, it takes the form of a standard two-step Generalized Method of Moment (GMM). In a high-dimensional context, *i.e.* when the number of covariates to account for in the choice of the weights is larger or proportional to the sample size, the weights are determined by a ℓ_1-minimization program in the same spirit as the Lasso. As a consequence, the resulting treatment effect estimator also suffers from the post-selection inference problem described above and needs to be "immunized". We do so drawing inspiration from the methodology developed by Chernozhukov et al. (2015). The proposed estimator is doubly robust, consistent and asymptotically normal uniformly over a large class of data-generating processes. We study its performance using Monte Carlo simulations and illustrate its application on the California tobacco control program originally studied in Abadie et al. (2010).

When I started working on this chapter, the use of the Lasso and its integration in a two-step framework where the second-step estimator is proven to be asymptotically Gaussian

was not as well-developed as they are now. Although some of the main contributions by Victor Chernozhukov and co-authors had already been published, the general framework as it appears, for example, in Chernozhukov et al. (2018a), was not available. From the perspective of the econometric theory, our work can be seen an application of this methodology before it was systematized.

2. Machine Learning in Empirical Economics

In this section, we adopt the potential outcome framework of Rubin (1974). Let D code for the treatment, so $D = 1$ for a treated individual and $D = 0$ otherwise. Let Y_1 and Y_0 be random variables representing potential outcomes under treatment and under no treatment, respectively. The effect of the treatment is $Y_1 - Y_0$. Let X be a random vector of observable individual characteristics.

2.1. State of the Art

Taking its roots in the post-selection inference problem, the integration of Machine Learning tools to the empirical economist's toolkit started with the Lasso. Likely spurred by the good theoretical understanding of this method and the sparsity of the resulting estimator, it has been studied in a number of settings and models relevant for practitioners, be it policy evaluation (Belloni et al., 2014b), instrumental variables (Belloni et al., 2012), panel data (Belloni et al., 2016), demand estimation (Chernozhukov et al., 2017a), discriminations (Bach et al., 2018), among others.

Many contributions in this literature were quick to highlight the potential benefits of these modern statistical tools for policy evaluation and causal inference, while acknowledging the difficulties posed by adapting them to achieve goals that are standard in empirical Economics and which are often broader than prediction (*e.g.* Athey, 2015; Athey and Imbens, 2016). At first, the perceived value-added of machine learning methods relied mostly in variable selection and high-quality estimation of high-dimensional nuisance parameters under an unconfoundedness assumption, *e.g.* Belloni et al. (2014b,a, 2017); Farrell (2015). Notice that while these methods had been applied in several empirical studies, they were not suited for inference until a few years ago. Recent contributions went beyond the Lasso to integrate other non-standard statistical tools of sufficiently high-quality to the empiricist's inference toolbox, such as random trees and forests, boosting, support vector machines, kernel methods, or neural networks (*e.g.* Chernozhukov et al., 2018a). Two ingredients played a key role in this breakthrough. First, the use of orthogonal scores ensures that the resulting treatment effect estimator is first-order insensitive to deviations from the true value of nuisance parameters (Chernozhukov et al., 2015; Chernozhukov et al., 2015; Chernozhukov et al., 2018). In other words, orthogonal scores mitigate the impact of replacing unknown nuisance parameters by machine learning estimators that are often not $\sqrt{n}$-consistent. Most of the time, such a strategy requires the estimation of more nuisance parameters, yielding the name of *double selection* (Belloni et al., 2014a) or *double machine learning* (Chernozhukov et al., 2017; Chernozhukov et al., 2018a). Second, sample-splitting appeared as a way to limit the proclivity of these tools to over-fit the data. Indeed, several papers (*e.g.* Athey and Wager, 2018) advocate for splitting the

data between an *auxiliary* sample where nuisance parameters are estimated and a *main* sample where the parameter of interest is estimated using out-of-sample predictions based on the predictors constructed on the auxiliary sample. Chernozhukov et al. (2017, 2018a) suggest that the role of the two samples be then switched and the estimators combined in order to prevent loss of efficiency, yielding the name *cross-fitting*. Finally, building on these elements, a recent contribution by Chernozhukov et al. (2018b) provides a framework to performance inference over some features of heterogeneous individual treatment effects while assuming very little about the performance of the chosen ML algorithm, hence the name of *generic machine learning*.

RCTs were, by design, less likely to benefit from the use of more complex econometric tools. Indeed, independence between the treatment variable and confounding factors immunizes RCT from a selection bias, granting them the status of 'gold standard' of scientific evidence among regularly studied designs (*e.g.* Abadie and Cattaneo, 2018). By leveraging the prediction performance of machine learning tools, inference regarding treatment effect in RCT can, however, benefit from an increase in precision (lower standard error). Furthermore, they offer a way of searching for treatment effect heterogeneity in the absence of a pre-analysis plan because they provide a flexible way to estimate the Conditional Average Treatment Effect (CATE), $E[Y_1 - Y_0|X]$, that is, the expected treatment effect conditional on some individual characteristics. For example, by their own natures, random trees and random forests partition the data to predict some outcome of interest (they find thresholds and interact variables in a data-driven fashion, Mullainathan and Spiess, 2017) and can be adapted for causal inference purposes (see the *causal trees* of Athey and Wager, 2018). Contrary to pre-analysis plans, they don't require to specify the dimensions along which the economist will search for heterogeneity beforehand while still not allowing for p-hacking if done correctly (Chernozhukov et al., 2018b). Indeed, pre-analysis plans can be costly because they are inflexible and end up wasting a lot of data points (Olken, 2015). In other words, machine learning 'lets the data speak' and allows to discover dimensions along which the treatment effect differs even if they previously were not suspected to matter.

So far, applications of these modern statistical methods to RCT have been scarce. Davis and Heller (2017b) and Davis and Heller (2017a) use causal forests to study the heterogeneity in effectiveness of a youth summer job program in reducing the probability of committing crimes and increasing the likelihood of attending school or being employed. Davis and Heller (2017a) successfully identify a sub-group for which the program increases employment while the effect of the program for employment is not statistically significant on average. This sub-group appears to differ from the youth usually targeted by these programs, thereby questioning the state of knowledge in this field.

2.2. Contribution

Chapter 4 of this thesis applies the Generic Machine Learning (Generic ML) framework developed by Chernozhukov et al. (2018b) to a job-training experiment with imperfect compliance in order to analyze the selection process of the applicants by the case-workers.

Put succinctly, the Generic ML framework allows to study heterogeneity in the effect of

the treatment as long as the econometrician can produce two quantities: a ML proxy of the treatment effect, *i.e.* a prediction of the individual treatment effect made using a machine learning algorithm, and an unbiased signal of the individual treatment effect, *i.e.* a random variable $\tilde{Y}$ such that $E[\tilde{Y}|X] = E[Y_1 - Y_0|X]$. They develop a test that allows to rule out the absence of heterogeneity and estimate some features of the distribution of the CATE. Contrary to standard RCTs with perfect compliance, a unbiased signal for the LATE does not exist and complicates the straightforward adaption of this convenient framework. We show that some features of the conditional LATE still can be recovered and propose other ways to analyze the selection of job-seekers by the case-workers.

We revisit a large-scale randomized experiment conducted in France in 2007-2008 that was created to evaluate the impact of an intensive job-search counseling program on employment outcomes. This experiment was specifically designed to compare a public and a private provisions of job counselling. Originally, Behaghel et al. (2014) found the public program to be twice as effective as the private program, a finding that they partially blame on the payment structure to which the private providers were subject. They suspect that it entailed two types of side effects: if the fixed part of the payment is large, private providers are likely to maximize enrollment intro the program and offer very little job counselling to keep the costs down (*parking*); if the conditional payment is relatively large, they are likely to enroll the candidates with the best labor market prospects and again, provide them with little job counselling (*cream-skimming*). On the other hand, the public arm of the program provided by the French Public Employment Service (PES), was not subject to financial incentives.

Using the Generic ML framework, we find evidence of heterogeneity in the baseline probability of finding a job and the treatment effect across individuals, in both the public and the private program. Preliminary results also show differential rates of enrollment across groups defined in terms of baseline outcome (outcome without the treatment) but not across groups defined in terms of individual treatment effect. For both the public and the private program, it suggests that case-workers targeted individuals who had lower labor market prospects.

2.3. Perspectives

While the integration of machine learning tools to the empiricist's toolbox has been studied fairly extensively and requires the application of procedures such as sample-splitting and the use of orthogonal scores, much remains to be done regarding the selection of the underlying ML algorithm. More specifically, two questions arise in this context: (i) is there any prior practical knowledge that can guide the econometrician in the choice of the ML algorithm? (ii) is there a data-driven procedure to select the best ML algorithm when the goal is inference regarding a particular parameter and not prediction?

Regarding the first question, the beginning of an answer can be found by looking at theoretical results for a particular ML estimator when they exist. For example, it is known that the Lasso requires sparsity in the true value of the parameter or at least *approximate sparsity* (Belloni et al., 2012). As a consequence, if the empirical economist suspects that only a few characteristics matter, using a Lasso makes sense. On the other

hand, if many characteristics are suspected to contribute little (the parameter is *dense*), a Ridge regression would be better. Similarly, when the regression function can be assumed to be piece-wise constant, a random tree seems fitting (Mullainathan and Spiess, 2017). However, on the one hand, most of these results pertain to a literature that can be quite remote from applied researchers and the clues found there may not be so practical; on the other hand, many ML algorithms, when they are applied in practice, differ quite substantially from their vanilla, theoretical counterpart – when they have been studied at all! Recently, several efforts have been made to provide guidance for application of these modern tools. Abadie and Kasy (2018) study different choices of regularization in a regression-type setting and provide guidance for the applied researcher as to when a particular type of regularization is likely to perform better than another. Knaus et al. (2018) study the performance of different machine learning estimators through Monte Carlo simulations when the goal is detection of heterogeneous treatment effects. However, theoretical contributions to this questions are likely to be of limited use while the answer may depend quite heavily on the application. I believe that practical applications and accumulated experience is key for further diffusion of these modern statistical tools.

Regarding the second question, notice that many empirical Economics questions are, in fact, prediction questions. For example, policy evaluation seeks to answer the question "what value would have taken the outcome had the policy not been implemented?", *i.e.* building a counterfactual can be viewed as a prediction task. The difficulty relies in the absence of *ground truth*, that is, of the observation of the outcome under no treatment for the treated units. Standard causal inference designs, such as unconfoundedness, allow for straightforward application of ML principles to form a counterfactual. Indeed, the assumption $E[Y_0|X, D = 1] = E[Y_0|X, D = 0]$ – once observed characteristics are controlled for, assignment to either treatment group does not help predicting the outcome without the treatment – allows to estimate the regression function over the untreated sample and apply it to the treated sample. All the usual ML arsenal such as cross-validation, etc. can be deployed on the untreated sample. Nevertheless, this adaptation of off-the-shelf ML tools is only partially satisfying. For instance, when looking for heterogeneous treatment effects, the target is the CATE, rather than each single regression function and gains are likely to be made by focusing on the right object, the CATE, see *e.g.* Künzel et al. (2019). Within that context of detecting heterogeneous effects, Chernozhukov et al. (2018b) develop measures of performance according to which pre-defined ML algorithms can be ranked when the goal is detecting heterogeneity. However, the realm of applications is limited by the requirement that the econometrician can construct an unbiased signal of the CATE directly from the data, a quantity that does not exists when the target is the conditional LATE. Furthermore, a model averaging perspective on this question maybe fruitful but has not been investigated, to the best of my knowledge.

Lastly, and partly disconnected from econometric theory, the emergence of ML and Artificial Intelligence (AI) opens the door to study machine-based decision making in Economics, *e.g.* Kleinberg et al. (2017, 2019).

3. Synthetic Control, High-Dimension and Selection of the Control Group

3.1. State of the Art

Since the original contributions of Abadie and Gardeazabal (2003); Abadie et al. (2010, 2015), synthetic control methods have often been applied to estimate the treatment effects of large-scale interventions (see, e.g., Kleven et al., 2013; Bohn et al., 2014a; Hackmann et al., 2015; Cunningham and Shah, 2018). Suppose we observe data for a unit affected by the treatment or intervention of interest, as well as data on a donor pool, that is, a set of untreated units that are available to approximate the outcome that would have been observed for the treated unit in the absence of the intervention. The idea behind synthetic controls is to match the unit exposed to the intervention of interest to a weighted average of the units in the donor pool that most closely resembles the characteristics of the treated unit before the intervention. Once a suitable synthetic control is selected, differences in outcomes between the treated unit and the synthetic control are taken as estimates of the effect of the treatment on the unit exposed to the intervention of interest. The simplicity of the idea behind synthetic control is probably one of the reasons why it has been considered "the most important innovation in the policy evaluation literature in the last fifteen years" by Athey and Imbens (2017) and has quickly garnered popularity in empirical research – see the practical guide by Abadie (2019).

Formally, let X_{treat} be the $p \times 1$ vector of pre-intervention characteristics for the treated unit. Let $X_1, \ldots, X_{n_0}$ be the same characteristics measured for the donor pool. In most applications, the p pre-intervention characteristics will contain pre-treatment outcomes (in which case $p = T_0$, the number of pre-treatment dates) but one might want to add other predictors of the outcome observed during the pre-treatment period. For a vector of dimension p, Z, and some real symmetric positive-definite matrix V of dimension $p \times p$, let $\|Z\|_V = Z'VZ$. The synthetic control solution $W^* = (W_1^*, \ldots, W_{n_0}^*)$ solves the program:

$$\min_{W \in \mathbb{R}^{n_0}} \left\| X_{treat} - \sum_{j=1}^{n_0} W_j X_j \right\|_V^2 \qquad \text{(SYNTH)}$$

$$\text{s.t.} \quad W_1 \geq 0, \ldots, W_{n_0} \geq 0, \qquad \text{(NON-NEGATIVITY)}$$

$$\sum_{j=1}^{n_0} W_j = 1. \qquad \text{(ADDING-UP)}$$

The resulting synthetic control estimator for the post-treatment dates $t = T_0 + 1, \ldots, T$ is the difference between the outcome for the treated and a convex combination of the outcomes of the untreated

$$\hat{\tau}_t = Y_{treat,t} - \sum_{j=1}^{n_0} W_j^* Y_{1,t}.$$

In recent years, many theoretical contributions came from studying synthetic controls in relation with panel data methods and factor models, notably the interactive fixed effect model of Bai (2009), *e.g.* Gobillon and Magnac (2016); Xu (2017). Several other contributions studied the bias of the synthetic control estimator when the pre-treatment

fit is imperfect (*e.g.* Fermand and Pinto, 2019) and proposed bias-corrected versions (*e.g.* Ben-Michael et al., 2019, Arkhangelsky et al., 2018). Inference for synthetic controls is also a prolific area of research. Indeed, the method is often applied in the context of long panel data where the standard inference procedure based on asymptotic theory may not provide a credible approximation of the true distribution of the estimator. For example, in Abadie et al. (2010), $T = 40$, $n_0 = 38$ and only one treated unit; in Acemoglu et al. (2016), $T \approx 300$, $n_0 = 513$ and a dozen treated units. Abadie et al. (2010) proposed using a type of cross sectional permutation-based inference closely related to Fisher exact tests (Imbens and Rubin, 2015, Chapter 5) where the treatment is re-assigned at random within units and the test statistics recomputed under that new assignment. Several contributions studied that procedure and the choice of tests statistics (*e.g.* Firpo and Possebom, 2018), while others adapted the so-called "conformal inference" framework where residuals are computed under the null hypothesis and permuted across the time dimension (Chernozhukov et al., 2017b).

While the connection with the high-dimensional/ML literature is not obvious at first glance, the synthetic control method illustrates well the recent developments in Econometrics, among which the integration of machine learning discussed previously is a trend. First of all, synthetic controls are especially powerful in a context where there is no common trend between the treatment and the control groups during the pre-treatment period. As such, the synthetic control method is a way to select a control group that matches the characteristics of the treated. In this context, the dimensionality of the problem is not related to p, the number of characteristics to be matched, but to n_0 the size of the donor pool. As a consequence, most of the time, synthetic controls require solving a high-dimensional problem, involving its specific set of difficulties. Furthermore, the synthetic control method, because it imposes little structure on the problem and offers a way to systematize the choice of the control group, shares the same "spirit" as machine learning. Several other contributions have also highlighted the connection between synthetic controls and matrix completion methods, *e.g.* Athey et al. (2017); Athey et al. (2019), introducing the possibility to use low-rank matrix factorization as a tool in this context. No wonder then, that machine learning journals publish research on synthetic control, *e.g.* Amjad et al. (2018).

3.2. Contribution

This thesis contains two chapters contributing to the synthetic control literature.

Chapter 2 solves the high-dimensionality question in synthetic controls by putting a parametric structure on the weights, thereby providing an alternative to the synthetic control method and connecting it to more standard econometric tools. The original synthetic control method has some limitations, in particular when applied to micro data, for which it was not initially intended. In such cases, the number of untreated units n_0 is typically larger than the dimension p of variables used to construct the synthetic units. Then, as soon as the treated unit falls into the convex hull defined by the untreated units, the synthetic control solution is not uniquely defined (see Chapter 3). Second, and still related to the fact that the method was not developed for micro data, there is yet, to

the best of our knowledge, no asymptotic theory available for synthetic control. This means in particular, that inference cannot be conducted in a standard way. A third issue is related to variable selection. The standard synthetic control method, as advocated in Abadie et al. (2010), not only minimizes the norm $\|.\|_V$ between the characteristics of the treated and those of its synthetic unit under constraints, but also optimizes over the weighting matrix V so as to obtain the best possible pre-treatment fit. This approach has been criticized for being unstable and yielding unreproducible results, see in particular Klößner et al. (2018). The proposed estimator addresses these issues.

Chapter 3 takes another perspective on the synthetic control method by viewing it as a type of matching estimator where the n_0 weights are assigned to the untreated units based on a program that maximizes the pre-treatment characteristics. The requirement that the weights are non-negative and sum to one provide some regularization but typically not enough to obtain a unique solution. In general, if the treated falls into the convex hull defined by the donor pool, the solution is not unique. In that case, the econometrician may turn the curse of dimensionality to his advantage by adding more variables to be matched on. However, this is neither always possible, nor desirable. To solve this problem, we introduce a penalization parameter that trades off pairwise matching discrepancies with respect to the characteristics of each unit in the synthetic control against matching discrepancies with respect to the characteristics of the synthetic control unit as a whole. This type of penalization is aimed to reduce interpolation biases by prioritizing inclusion in the synthetic control of units that are close to the treated in the space of matching variables. Moreover, we show that as long as the penalization parameter is positive, the generalized synthetic control estimator is unique and sparse. If the value of the penalization parameter is close to zero, our procedure selects the synthetic control that minimizes the sum of pairwise matching discrepancies (among the synthetic controls that best reproduce the characteristics of the treated units). If the value of the penalization parameter is large, our estimator coincides with the pair-matching estimator. We study both the geometric properties of the penalized synthetic control solution and the large-sample properties of the resulting estimator, and propose data driven choices of the penalization parameter. We also propose a bias-corrected version of the synthetic control estimator. We complete this chapter by Monte Carlo simulations and two empirical studies.

3.3. Perspectives

The use of the synthetic control method among empirical researchers is likely to keep growing and along with it, the development of more data-driven methods to choose the comparison group in a transparent manner.

A few particular aspects of the original method remained to be explored in a theoretical fashion. For example, the original paper by Abadie et al. (2010) replaced (SYNTH) by the objective:

$$\min_{W \in \mathbb{R}^{n_0}} \sum_{k=1}^{p} \mathbf{v_k} \left(X_{treat,k} - \sum_{j=1}^{n_0} W_j X_{j,k} \right)^2 ,$$

where the positive weights $\mathbf{v_1}, \ldots, \mathbf{v_p}$ reflect the importance given to each predictor of the outcome in the minimization problem. The authors also proposed a data-driven way of choosing these weights so as to minimize the discrepancy between the synthetic unit and the treated unit in the outcome during the pre-treatment period. However, Klößner et al. (2018) show that the solution found is not uniquely defined which can be a problem for robustness and reproducibility of the results. When the choice of weights $\mathbf{v_1}, \ldots, \mathbf{v_p}$ does and does not matter and how to regularize this kind of cross-validation procedure is unknown at the moment.

Chapter 2

A Parametric Alternative to the Synthetic Control Method with Many Covariates

Joint work with Marianne Bléhaut, Xavier D'Haultfœuille and Alexandre Tsybakov.

Summary

The synthetic control method developed by Abadie et al. (2010) is a an econometric tool to evaluate causal effects when a few units are treated. While initially aimed at evaluating the effect of large-scale macroeconomic changes with very few available control units, it has increasingly been used in place of more well-known microeconometric tools in a broad range of applications, but its properties in this context are unknown. This paper proposes a parametric generalization of the synthetic control, which is developed both in the usual asymptotic framework and in the high-dimensional one. The proposed estimator is doubly robust, consistent and asymptotically normal uniformly over a large class of data-generating processes. It is also immunized against first-step selection mistakes. We illustrate these properties using Monte Carlo simulations and applications to both standard and potentially high-dimensional settings, and offer a comparison with the synthetic control method.

1. Introduction

The original synthetic control method developed by Abadie and Gardeazabal (2003); Abadie et al. (2010, 2015) is an econometric tool to quantify the effects of a policy change that affects one or very few aggregate units, using aggregate-level data. The idea is to construct a counterfactual treated unit using a convex combination of non-treated units, the "synthetic control unit", that closely recreates the characteristics of the treated. The weight given to each control unit are computed by minimizing the discrepancy between the treated and the synthetic unit in the mean of predictors of the outcome of interest. The synthetic control method has been used to evaluate causal impacts in a wide range of applications such as terrorism, civil wars and social unrest (Acemoglu et al., 2016), political and monetary unions (Abadie et al., 2015, Wassmann, 2015), minimum wage (Dube and Zipperer, 2015, Addison et al., 2014), health (Bilgel and Galle, 2015), fiscal policies (Dietrichson and Ellegård, 2015), geographical and regional policies (Gobillon and Magnac, 2016), immigration policy (Bohn et al., 2014b), international trade (Nannicini and Billmeier, 2011) and many more. While initially aimed at evaluating the effect of large-scale macroeconomic changes with very few available units of comparison, most of the time these units being states or regions, the synthetic control method has increasingly been used in place of more well-known microeconometric tools. Contrasting with these standard approaches, the theory behind the synthetic control estimator has not been fully built yet, especially when the number of control units tends to infinity.

This paper proposes an alternative to the synthetic method by using a parametric form for the weight given to each control unit. In the small-dimensional case where the number of observations is much larger than the number of covariates, our approach amounts to a two-step GMM estimator, where the parameters governing the synthetic weights are computed in a first step so that the reweighted control group matches some features of the treated. A key result of the paper is the double robustness of the estimator, as defined by Bang and Robins (2005). Under that property, misspecifications in the synthetic control weights do not prevent valid inference if the outcome regression function is linear for the control group. This approach is also extended to the high-dimensional case where the number of covariates is proportional or larger than the number of observations and to cases where variable selection is performed. This extension makes the proposed estimator suitable for comparative case studies and macroeconomic applications. Here, the double robustness property helps constructing an estimator which is *immunized* against first-step selection mistakes in the sense defined by Chernozhukov et al. (2015); Chernozhukov et al. (2018a). In both cases, it is consistent and asymptotically normal uniformly over a large class data-generating processes. Consequently, we develop inference based on asymptotic approximation, linking the synthetic control method with more standard microeconometric tools.

The present paper builds mainly along two lines of the treatment effect literature. The first one is the literature related to propensity score weighting and covariate balancing propensity scores. Several recent efforts have been made to include balance between covariates as an explicit objective for estimation with or without relation to the propensity score (*e.g.* Hainmueller (2012); Graham et al. (2012)). Recently, Imai and Ratkovic

(2014) integrated propensity score estimation and covariate balancing in the same framework. Their covariate balancing propensity score method is estimated with GMM and yields more robust estimates than standard propensity score-related methods. Indeed, they show that this method is less impacted by potential misspecifications and retains the theoretical properties of GMM estimators. Our Theorem 2.1 gives a theoretical basis to support these empirical findings. It is to be noted that the covariate balancing idea is related to the *calibration on margins* method used in survey sampling, see for example Deville et al. (1993).

It also partakes in the econometric literature that addresses variable selection, and more generally the use of machine learning tools, when estimating a treatment effect, especially but not exclusively in a high-dimensional framework. The lack of uniformity for inference after a selection step has been raised in a series of papers by Leeb and Pötscher (2005, 2008a,b), echoing earlier papers by Leamer (1983) who put into question the credibility of many empirical policy evaluation results. One recent innovative solution proposed to circumvent this post-selection conundrum is the use of double-selection procedures Belloni and Chernozhukov (2013); Farrell (2015); Chernozhukov et al. (2015); Chernozhukov et al. (2018a). For example, Belloni et al. (2014a,b) highlight the dangers of selecting controls exclusively in their relation to the outcome and propose a three-step procedure that helps selecting more controls and guards against omitted variable biases much more than a simple "post-single-selection" estimator, as it is usually done by selecting covariates based on either their relation with the outcome or with the treatment variable, but rarely both. Farrell (2015) extends this approach by allowing for heterogeneous treatment effects, proposing an estimator that is robust to either model selection mistakes in propensity scores or in outcome regression. In addition, he deals explicitly with a discrete treatment that is a more common setting in the policy evaluation literature. Chernozhukov et al. (2015, 2018a) have theorized this approach by showing how using moments that are first-order-insensitive to the selection step help *immunizing* the inference against selection mistakes, or more generally against estimators that are not $\sqrt{n}$-consistent. A different path to deal with the problem of propensity score specification has been followed by Kitagawa and Muris (2016) using the Focused Information Criterion (FIC) of Claeskens and Hjort (2003), but it does not explicitly accommodate for a high-dimensional nuisance parameter and assumes that the researcher knows the true model.

The paper is organized as follows. Section 2 introduces our estimator and states its properties in a standard low-dimensional setting. Section 3 extends the previous section to the high-dimensional case and studies its asymptotic properties. Section 4 illustrates the good inference properties of the estimator in a Monte Carlo experiment. Section 5 revisits LaLonde (1986)'s dataset to compare our procedure with other high-dimensional econometric tools and the effect of the large-scale tobacco control program of Abadie et al. (2010) for a comparison with synthetic control. The appendix gathers the proofs.

2. A Parametric Alternative to Synthetic Control

2.1. Covariate Balancing Weights and Double Robustness

We are interested in the effect of a binary treatment, coded by $D = 1$ for the treated and $D = 0$ for the non-treated. We let Y_0 and Y_1 denote the potential outcome under no treatment and under the treatment, respectively. The observed outcome is then $Y = DY_1 + (1 - D)Y_0$. We also observe a random vector $X \in \mathbb{R}^p$ of pre-treatment characteristics. The quantity of interest is the Average Treatment Effect on the Treated (ATET) defined a:

$$\theta_0 = \mathbb{E}[Y_1 - Y_0|D = 1].$$

Since no individual is observed in both treatment states, identification of the counterfactual $\mathbb{E}[Y_0|D = 1]$ is achieved through the following two ubiquitous conditions.

Assumption 2.1 (Nested Support) $\mathbb{P}[D = 1|X] < 1$ *almost surely and* $\pi := \mathbb{P}[D = 1] \in (0, 1)$.

Assumption 2.2 (Mean Independence) $\mathbb{E}[Y_0|X, D = 1] = \mathbb{E}[Y_0|X, D = 0]$.

Assumption 2.1, a version of the usual common support condition, requires that there exist control units for any possible value of the covariates in the population. Since the ATET is the parameter of interest, we are never reconstructing a counterfactual for control units so $\mathbb{P}[D = 1|X] > 0$ is not required. Assumption 2.2 states that conditional on a set of observed confounding factors, the expected potential outcome under no treatment is the same for treated and control individuals. This assumption is a weaker form of the classical conditional independence assumption : $(Y_0, Y_1) \perp\!\!\!\perp D|X$.

As in most of the time in policy evaluation settings, the counterfactual is identified and estimated as a weighted average of non-treated unit outcomes:

$$\theta_0 = \mathbb{E}[Y_1|D = 1] - \mathbb{E}[WY_0|D = 0], \tag{2.1}$$

where W is a random variable. Popular choices for the weights are the following:

1. Linear regression: $W = \mathbb{E}[DX']\mathbb{E}[(1 - D)XX']^{-1}X$, also referred to as the Oaxaca-Blinder estimator Kline (2011),
2. Propensity score: $W = P[D = 1|X]/(1 - P[D = 1|X])$,
3. Matching: see Smith and Todd (2005) for more details,
4. Synthetic controls: see Abadie et al. (2010).

This paper proposes another choice of weight W which can be seen as a particular solution of the synthetic control. Formally, we look for weights W that (i) satisfy a balancing

condition as in the synthetic control method, are (ii) positive and (iii) function of the covariates. The first condition writes:

$$\mathbb{E}[DX] = \mathbb{E}[W(1-D)X]. \tag{2.2}$$

Up to a proportional constant, this is equivalent to $\mathbb{E}[X|D=1] = \mathbb{E}[WX|D=0]$. This condition means that W balances the first moment of the observed covariates between the treated and the control group. The definition of the observable covariates X is left to the econometrician and can include transformation of the original covariates so as to match more features of their distribution. The idea behind such weights relies on the idea of "covariate balancing" as in *e.g.* Imai and Ratkovic (2014). The following lemma shows that under Assumption 2.1, weights satisfying the balancing condition always exist.

Lemma 2.1 (Balancing Weights) *If Assumption 2.1 holds, the propensity score weight* $W_0 := \mathbb{P}[D=1|X]/(1-\mathbb{P}[D=1|X])$ *satisfies the balancing condition* (2.2).

It is straightforward to verify by plugging this expression in equation (2.2) and using the law of iterated expectations. Note that the linear regression weight $W = \mathbb{E}[DX']\mathbb{E}[(1-D)XX']^{-1}X$ also verifies the balancing condition but can be negative. The lemma suggests estimating a binary choice model to obtain $\mathbb{P}[D=1|X]$ and estimate weights W_0 as a first step, and plugging them to estimate θ_0 in a second step. However, an inconsistent estimate of the propensity score leads to an inconsistent estimator of θ_0 and does not guarantee that the implied weights will achieve covariate balancing. Finally, estimation of a propensity score can be problematic when there are very few treated units. For these reasons, we consider instead an estimation directly based on balancing equations:

$$\mathbb{E}\left[(D-(1-D)W_0)\,X\right] = 0. \tag{2.3}$$

An important advantage of this approach over the usual one based on the propensity score estimation through maximum likelihood is its double-robustness (for a definition, see, e.g., Bang and Robins, 2005). Indeed, let W_1 denote the weights identified by (2.3) and a misspecified model on the propensity score. Because the balancing equations (2.3) still hold for W_1, the estimated treatment effect will still be consistent provided that $\mathbb{E}[Y_0|X]$ is linear in X. The formal result is provided in Theorem 2.1 below.

We consider a parametric estimator of W_0. Suppose that $P[D=1|X] = G(X'\beta_0)$ for some unknown $\beta_0 \in \mathbb{R}^p$ and some known, strictly increasing cumulative distribution function G. Then $W_0 = h(X'\beta_0)$ with $h = G/(1-G)$ and β_0 is identified by (2.3). h is a positive increasing function, meaning that its primitive H is convex and its derivative (if it exists) is positive. A classical example of h would be $h = \exp$, corresponding to a logistic distribution for G. In such an example, $h = h' = H$. In any case, the convexity of H implies that β_0 is the solution of the strictly convex program:

$$\beta_0 = \arg\min_{\beta} \mathbb{E}\left[(1-D)H(X'\beta) - DX'\beta\right]. \tag{2.4}$$

Note that this program is well-defined, whether or not $P[D=1|X] = G(X'\beta_0)$.

We are now ready to state the main identification theorem that justifies the use of the ATET estimand of equation (2.1):

Theorem 2.1 (Double Robustness) *Suppose that Assumptions 2.1-2.2 hold and let β_0 defined by equation (2.4) for some positive, strictly increasing convex function H. Then, for any $\mu \in \mathbb{R}^p$, θ_0 satisfies*

$$\theta_0 = \frac{1}{\mathbb{E}(D)}\mathbb{E}\left[(D - (1-D)h(X'\beta_0))\,(Y - X'\mu)\right], \tag{2.5}$$

in two cases:

1. *the regression function under no treatment is linear, i.e. there exists $\mu_0 \in \mathbb{R}^p$ such that $\mathbb{E}[Y_0|X] = X'\mu_0$, or*

2. *the propensity score is given by $P[D = 1|X] = G(X'\beta_0)$, with $G = h/(1 + h)$.*

Theorem 2.1 highlights the double-robustness property of using an estimate of the propensity score based on the balancing approach. This result is similar to the one obtained by Kline (2011) for the Oaxaca-Blinder estimator, but his requires the propensity score to follow specifically a log-logistic model in the propensity-score-well-specified case. So Theorem 2.1 is more general. At this stage, μ in equation (2.5) does not play any role and could be zero. However, we will see below that choosing carefully μ is important in the high-dimensional case to obtain an "immunized" estimator of θ_0.

2.2. Asymptotic Properties in Low-Dimension

Consider an asymptotic regime within which the dimension p of the covariates is fixed, while the sample size n tends to infinity. An estimator of β_0 is obtained by taking the empirical counterpart of (2.4):

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}(1 - D_i)H(X_i'\beta) - D_i X_i'\beta. \tag{2.6}$$

Including an intercept among the X is strongly advised as it ensures that estimated weights sum to one. This estimator is plugged in the empirical counterpart of (2.5) to estimate θ_0:

$$\widetilde{\theta} := \frac{1}{\frac{1}{n}\sum_{i=1}^{n} D_i}\left(\frac{1}{n}\sum_{i=1}^{n}[D_i - (1 - D_i)h(X_i'\widehat{\beta})]Y_i\right).$$

Denote the estimating moment for θ_0 by $g(Z, \theta, \beta, \mu) := [D-(1-D)h(X'\beta)][Y-X'\mu]-D\theta$. In the low-dimensional case, $\widetilde{\theta}$ is such that

$$\frac{1}{n}\sum_{i=1}^{n} g(Z_i, \widetilde{\theta}, \hat{\beta}, 0) = 0$$

This estimator is a two-step GMM which is consistent and asymptotically normal with variance given by

$$\sigma^2 := \mathbb{E}\left[g(Z, \theta_0, \beta_0, \mu_0)^2\right] / \mathbb{E}(D)^2,$$

where $\mu_0 := \mathbb{E}[h'(X'\beta_0)XX'|D=0]^{-1}\mathbb{E}[h'(X'\beta_0)XY|D=0]$, under mild regularity conditions, see Section 6 in Newey and McFadden (1994). The quantity μ_0, appearing in the variance, is the coefficient of the weighted population regression of Y on X for the control group. The next section will use this observation to adapt the estimation in the high-dimensional case.

3. High-Dimensional Covariates and Post-Selection Inference

3.1. Regularized Estimation

In practice, the empirical researcher can be faced with a high-dimensional set of covariates in several situations:

1. In some applications the researcher is faced with a large dataset in the sense that many covariates are to be considered with respect to the relatively small sample size. It is a natural setting that often occurs in macroeconomic problems. For example, in the Tobacco control program application by Abadie et al. (2010) the control group size is limited due to the fact that the observational unit is the state but many pre-treatment outcomes are included among the covariates. Section 5 revisits this example.

2. Sometimes the researcher also wants to consider a flexible form for the weights and wants to include transformations of the covariates. This arises for example when categorical variables are interacted with other categorical variables or with continuous variables, or when a discrete variable such as the number of schooling years is broken down into binary variables to have a very flexible non-linear effect. This case can be labeled as "non-parametric".

3. More specifically in our estimation strategy, one may want not only to balance the first moments of the distribution of the covariates but also the second moments, the covariances, the third moments and so on to make the distribution more similar between the treated and the control group. In this case, a high-dimension setting appears to be desirable.

An inherent element of the high-dimensional literature is the notion of *sparsity*, *i.e.* the assumption that although we consider many variables, only a small number of elements in the vector of parameter is different from zero. This assumption amounts to recasting the problem in a variable selection framework where a good estimator should be able to correctly select the relevant variables or approximate the quantities of interest and be consistent at a rate close to $\sqrt{n}$, only paying a price proportional to the number of non-zero elements. A less restrictive concept has been introduced by Belloni et al. (2012). Called *approximate sparsity*, it assumes that the high-dimensional parameter can be decomposed

into a sparse component, which has many zero entries and some large entries in absolute value, and a small component for which all entries are small and decaying towards zero without never exactly being zero. It has been shown that in both contexts, Lasso-type estimators can provide a good approximation of the relevant quantities that are subject to a sparse structure, be it finite or infinite dimensional parameters. Consequently, consider the program (2.6), regularized by penalizing the ℓ_1-norm of β:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (1 - D_i) H(X_i'\beta) - D_i X_i'\beta + \lambda_d \sum_{j=1}^{p} \psi_{d,j} |\beta_j|, \tag{2.7}$$

where $\lambda_d > 0$ is an overall penalty parameter set to dominate the noise that stems from the gradient of the function and $\{\psi_{d,j}\}_{j=1,\dots,p}$ are covariate specific penalty loadings set as to insure good asymptotic properties. The penalty loadings are estimated using the algorithm presented in the appendix. For empirical applications, we advise not to penalize the intercept in order to obtain final weights that sum to one by construction.

The form of this minimization program is one of the main contributions of this paper to the existing literature on high-dimensional models. On the one hand, this program targets covariate balancing as the main objective because equating the derivative of the loss function to zero yields a balancing condition as in equation (2.3). On the other hand, this objective function includes a term that penalizes the complexity of the model. Such an objective function borrows from the Lasso estimator of Tibshirani (1996), further studied and generalized most notably in Candes and Tao (2007); Van de Geer (2008); Bickel et al. (2009). It has been specifically studied in the econometric literature by Belloni et al. (2012); Belloni and Chernozhukov (2013). This type of penalization offers multiple advantages: it regularizes the program so as to make it solvable contrary to a non-penalized GMM estimator, it yields strict sparsity in the sense that some elements of the estimated coefficients will be set exactly to zero if the penalty is large enough contrary to an ℓ_2-penalization, it is computationally feasible because it gives rise to a convex program contrary to an ℓ_0-penalization. The use of covariate-specific penalty loadings borrows from the approach of Belloni et al. (2012) that adapts the Lasso to the non-Gaussian, non-homoscedastic case. The drawback of penalizing by the ℓ_1-norm is the bias that it induces in the estimation of the coefficients. To remove it, a popular solution given in the Lasso-related econometric literature is the use of a Post-Lasso estimator. Such an estimator would use a second step where variables corresponding to non-zero elements of $\hat{\beta}$ are kept in the model and the others are discarded. Then estimation is done a second time using only these variables and no penalization to compute the Post-Lasso solution (see Belloni and Chernozhukov, 2013). Our strategy allows this estimator to be used, although we do not pursue this avenue here.

The estimator of β_0 above will be consistent as n tends to infinity under classical assumptions used for the Lasso with quadratic loss, see Theorem 2.3 below. As suggested above, we could then consider the plug-in estimator for the ATT, based on Equation (2.5) with $\mu = 0$:

$$\widetilde{\theta} = \frac{1}{\sum_{i=1}^{n} D_i} \sum_{i=1}^{n} [D_i - (1 - D_i) h(X_i'\widehat{\beta})] Y_i.$$

We refer to this estimator as the *naive plug-in estimator*. Notice that the Lasso estimator of the nuisance parameter β_0 is not asymptotically Normal. Intuitively, it cannot be the case since there is a non-zero probability that an entry of $\hat{\beta}$ is equal to zero due to the ℓ_1-penalization. Because a high-dimensional setting requires an asymptotic framework where p grows with n, the naive plug-in estimator will suffer from a regularization bias and may not be asymptotically Normal, as illustrated for example in Belloni et al. (2014b); Chernozhukov et al. (2015, 2018a).

3.2. Immunized Estimation

Following Chernozhukov et al. (2015, 2018a), consider an immunized estimator that is first-order insensitive to $\widehat{\beta}$. This estimator will be asymptotically normal with a very simple asymptotic variance that does not depend on the properties of the first-step estimator. The idea is to choose a μ in (2.5) so that the derivative of this moment with respect to β is zero when taken at (θ_0, β_0). This holds for $\mu = \mu_0$, where μ_0 satisfies

$$\mathbb{E}\left[(1-D)h'(X_i'\beta_0)(Y - X'\mu_0)X\right] = 0.$$

Notice that since h is a strictly increasing function of its argument, h' will be positive. Recognizing the first order condition of a least-squares program, μ_0 can be obtained as the coefficient of a weighted regression of Y on X for the control group:

$$\mu_0 = \arg\min_{\mu} \mathbb{E}\left[(1-D)h'(X'\beta_0)(Y - X'\mu)^2\right]. \tag{2.8}$$

Note that we have to estimate μ_0 which is also of dimension p. Notice that by construction the derivative of the moment condition (2.5) with respect to (β, μ) is equal to zero at the true values (β_0, μ_0) so we are not introducing another source of nuisance in the estimation. Consider once again a Lasso-type estimator for μ_0:

$$\hat{\mu} = \arg\min_{\mu} \frac{1}{n}\sum_{i=1}^{n}(1-D_i)h'\left(X_i'\hat{\beta}\right)\left(Y_i - X_i'\mu\right)^2 + \lambda_y \sum_{j=1}^{p}\psi_{y,j}|\mu_j|. \tag{2.9}$$

As previously, $\lambda_y > 0$ is an overall penalty parameter set to dominate the noise that stems from the gradient of the function, and $\{\psi_{y,j}\}_{j=1,\ldots,p}$ are covariate-specific penalty loadings. Finally, the immunized ATT estimator is defined as

$$\begin{aligned}\widehat{\theta} &:= \frac{1}{\sum_{i=1}^{n} D_i}\sum_{i=1}^{n}\left(D_i - (1-D_i)h(X_i'\hat{\beta})\right)\left(Y_i - X_i'\hat{\mu}\right) \\ &= \underbrace{\tilde{\theta}}_{\text{Naive Plug-In}} - \underbrace{\left[\frac{1}{n_1}\sum_{i:D_i=1}^{n} X_i - \frac{1}{n_1}\sum_{i:D_i=0}^{n} h(X_i'\hat{\beta})X_i\right]'\hat{\mu}}_{\text{Correction = Imbalance} \times \text{ Outcome-related } X}.\end{aligned}$$

Intuitively, the immunized moment corrects the naive plug-in estimate in the case where

the balancing program has "missed" a covariate which appears to be very important to predict the outcome. This result has a flavor of Frish-Waugh-Lowell partialling-out procedure for model selection as put under the spotlights most notably by Belloni et al. (2014a) and further theorized in Chernozhukov et al. (2015). Indeed, the estimating moment (2.5) for θ_0 can be re-written so as to highlight the partialling out of X from both Y and D:

$$\mathbb{E}\left(\underbrace{[D-(1-D)h(X'\beta_0)]}_{\text{Residual Imbalance}}\underbrace{[Y-X'\mu_0]}_{\text{Residual from Regression}}\right)=\mathbb{E}\left(D\theta_0\right)$$

Here, the effect of X is taken out from Y in a linear fashion, while the effect of X on D is taken out by re-weighting the control group so as to yield the same mean for X.

To summarize, the estimator in the high-dimensional case comprises the three following steps. Each step is simple to obtain as it involves at most to minimize a convex (and in general strictly convex) function:

1. *(Balancing step)* For a given level of penalty λ_d and positive covariate-specific penalty loadings $\{\psi_{d,j}\}_{j=1}^p$ solve the following:

$$\hat{\beta}=\arg\min_{\beta}\frac{1}{n}\sum_{i=1}^{n}(1-D_i)H(X_i'\beta)-D_iX_i'\beta+\lambda_d\sum_{j=1}^{p}\psi_{d,j}|\beta_j|, \tag{2.10}$$

2. *(Immunization step)* For a given level of penalty λ_y and covariate-specific penalty loadings $\{\psi_{y,j}\}_{j=1}^p$ solve the following, using $\hat{\beta}$ estimated in the previous step:

$$\hat{\mu}=\arg\min_{\mu}\frac{1}{n}\sum_{i=1}^{n}(1-D_i)h'(X_i'\hat{\beta})\left(Y_i-X_i'\mu\right)^2+\lambda_y\sum_{j=1}^{p}\psi_{y,j}|\mu_j|, \tag{2.11}$$

3. *(ATT estimation)* Estimate the ATT using the immunized moment estimator:

$$\widehat{\theta}=\frac{1}{\sum_{i=1}^{n}D_i}\sum_{i=1}^{n}\left[D_i-(1-D_i)h(X_i'\hat{\beta})\right](Y_i-X_i'\hat{\mu}). \tag{2.12}$$

We will refer to the estimator $\widehat{\theta}$ as the *immunized* estimator.

3.3. Asymptotic Properties

The current framework poses several challenges to achieving inference that would be uniform on a large class of DGP. Firstly, X is of high-dimension, since we allows $p > n$ and p can grow with n under the conditions stated in Assumption 2.3. Secondly, the ATT estimation is affected by estimation of nuisance parameters β_0 and μ_0 and we wish

to neutralize their influence. Finally and closely related to high-dimensional statistical problems, the ℓ_1-penalized estimators we use for β_0 and μ_0 are not conventional. The estimator of β_0 relies on a convex but potentially non-Lipschitz loss function, contrary to cases considered by Van de Geer (2008). The estimation of μ_0 is close to a standard Lasso except that it relies on weights depending on $\hat{\beta}$.

For the sake of brevity, we still denote the observed data by $Z_i := (Y_i, D_i, X_i)$ and let $\eta_0 := (\beta_0, \mu_0)$ denote the vector gathering the two nuisance parameters. Also denote the estimating moment for θ_0 by $g(Z, \theta, \eta) = g(Z, \theta, \beta, \mu)$. We recall that the true values (θ_0, η_0) satisfy

$$\mathbb{E}g(Z, \theta_0, \eta_0) = 0. \tag{2.13}$$

Before introducing our assumptions, we introduce additional notations. Hereafter, $a \lesssim b$ means that $a \leq cb$ for some constant $c > 0$ independent of the sample size n. Φ and Φ^{-1} are the distribution and quantile functions of a standard Normal random variable. $\mathbb{E}_n(.)$ denotes the average over index i, $n^{-1}\sum_{i=1}^n(.)$. For a vector $\delta \in \mathbb{R}^p$, $\|\delta\|_0 := \text{Card}\{1 \leq j \leq p, \delta_j \neq 0\}$, $\|\delta\|_1 := \sum_{j=1}^p |\delta_j|$, $\|\delta\|_2 := \sqrt{\sum_{j=1}^p \delta_j^2}$, $\|\delta\|_\infty := \max_{j=1,...,p} |\delta_j|$.

Assumption 2.3 (Sparsity and Dimension Restrictions) *(i) The nuisance parameter $\eta_0 := (\beta_0', \mu_0')'$ is sparse in the following sense:*

$$\|\beta_0\|_0 \leq s_\beta, \ \|\mu_0\|_0 \leq s_\mu.$$

(ii) Growth condition: $\log(p)(s_\beta + s_\mu)/\sqrt{n} \to 0$,
(iii) $s_\beta \sim s_\mu$.

Assumption 2.4 (Conditions on the Design) *Consider a sequence $\{\boldsymbol{P_n}\}_{n\in\mathcal{N}}$ of sets of probability measures such that for each sequence $\{\mathbb{P}_n\}_{n\in\mathcal{N}} \in \{\boldsymbol{P_n}\}_{n\in\mathcal{N}}$, the following hold. (Y_i, X_i, X_i) are i.i.d. random vectors such that:*

$\|X_i\|_\infty \leq K_n$, *a.s.*,
$\|X_i'\beta_0\|_\infty \leq K_n'$, *a.s.*,
$\|Y_i - X_i'\mu_0\|_\infty \leq K_n''$, *a.s.*,
$\mathbb{P}(D_i = 1) \in (0, 1)$,
h'' is Lipschitz on any compact subset of $\mathbb{R}$,
either h' is bounded away from zero or $\|\beta_0\|_1$ is bounded,
$\liminf_n \min\left\{\mathbb{E}\left((Y_{1i} - X_i'\mu_0 - \theta_0)^2 | D_i = 1\right), \mathbb{E}\left(h(X_i'\beta_0)^2(Y_{0i} - X_i'\mu_0)^2 | D_i = 0\right)\right\} > 0.$

The following condition holds with $C_n \in \{K_n, K_n', K_n''\}$:

$$C_n^2 s \log(n)^2 \log(s\log(n))^2 \log(p \vee n) = o(nc_\kappa^4/c_\phi).$$

Define $\Sigma := \mathbb{E}((1 - D)XX')$, the theoretical Gram matrix on the control group. For a

non-empty subset $S \subset \{1, ..., p\}$ and $\alpha > 0$, define the set:

$$\mathcal{C}[S, \alpha] := \{v \in \mathbb{R}^p : \|v_{S^C}\|_1 \leq \alpha \|v_S\|_1, v \neq 0\} \tag{2.14}$$

Assumption 2.5 (Conditions on the Gram matrix for the control group) *For integers $s := \max(s_\beta, s_\mu)$, p such that $1 \leq s \leq p/2$,$m \geq s$, $s + m \leq p$, a vector $\delta \in \mathbb{R}^p$ and a set of indices S with $|S|_0 \leq s$, denote by S_1 the subset of $\{1, ..., p\}$ corresponding to the m largest in absolute value coordinate of δ outside of S and define $S_{01} = S \cup S_1$. For $\alpha = c_0 c_\psi$,*

$$\kappa_\alpha^2(\Sigma) := \min_{\substack{S \subset \{1,...,p\} \\ |S|_0 \leq s}} \min_{\delta \in \mathcal{C}[S,\alpha]} \frac{\delta' \Sigma \delta}{\|\delta_{S_{01}}\|_2^2} > 0.$$

In order to analyze the sparsity of the estimator, a bound on the maximal eigenvalue is needed. It exist constants c_κ and c_ϕ such that:

$$0 < c_\kappa^2 \leq \min_{\|\delta\|_0 \leq s \log n} \frac{\delta' \Sigma \delta}{\|\delta\|_2} \leq \max_{\|\delta\|_0 \leq s \log n} \frac{\delta' \Sigma \delta}{\|\delta\|_2} \leq c_\phi < \infty.$$

Define the random variable V_i such that

$$V_i := \max \left\{ h''(X_i'\beta_0), h''(X_i'\beta_0)(Y_i - X_i'\mu_0), |h'(X_i'\beta_0)| \right\}.$$

There exist a finite fixed constant c_ϕ' such that:

$$\max_{\|\delta\|_0 \leq s \log n} \frac{\delta' \mathbb{E}\left((1 - D_i) V_i X_i X_i'\right) \delta}{\|\delta\|_2^2} \leq c_\phi'.$$

Moreover, there exists a constant $c_\Sigma > 1$ such that for all $v \in \mathbb{R}^p$,

$$\sqrt{\mathbb{E}[(v'(1 - D_i) X_i X_i' v)^2]} \leq c_\Sigma v' \Sigma v.$$

Assumption 2.6 (Penalty Loadings) *Let $c > 1$, $\gamma \lesssim \log(p \vee n)$ and β_0 denote the true coefficient.*[1] *The ideal penalty loadings for estimation of β_0 are given by:*

$$\lambda^d := c\Phi^{-1}(1 - \gamma/2p)/\sqrt{n}$$

$$\psi_{d,j} := \sqrt{\frac{1}{n} \sum_{i=1}^n \left[(1 - D_i) h(X_i'\beta_0) - D_i\right]^2 X_{i,j}^2}, \text{ for } j = 1, ..., p$$

The ideal penalty loadings for estimation of μ_0 are given by:

$$\lambda^y := 2c\Phi^{-1}(1 - \gamma/2p)/\sqrt{n}$$

$$\psi_{y,j} := \sqrt{\frac{1}{n} \sum_{i=1}^n (1 - D_i) h'(X_i'\beta_0)^2 \left[Y_i - X_i'\mu_0\right]^2 X_{i,j}^2}, \text{ for } j = 1, ..., p$$

[1] Belloni et al. (2012) set $\gamma := 0.1/\log(p \vee n)$ and $c := 1.1$.

Moreover,

$$\bar{\psi} := \max_{1\leq j\leq p} \max(\psi_{d,j}, \psi_{y,j}) < \infty, \; \underline{\psi} := \min_{1\leq j\leq p} \min(\psi_{d,j}, \psi_{y,j}) < \infty. \tag{2.15}$$

Finally, $c_\psi := \bar{\psi}/\underline{\psi}$.

The following theorem constitutes the main asymptotic result of the paper.

Theorem 2.2 (Asymptotic Normality of the Immunized Estimator) *Consider a sequence* $\{\boldsymbol{P_n}\}$ *of sets of probability measures such that for each sequence* $\{\mathbb{P}_n\} \in \{\boldsymbol{P_n}\}$ *the assumptions 2.3 - 2.6 hold. The immunized estimator* $\hat{\theta}$ *defined in equation* (2.12) *verifies:*

$$\hat{\sigma}^{-1}\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0,1), \text{ as } n \to \infty,$$

where $\hat{\sigma}^2 := \mathbb{E}_n\left[g(Z_i, \hat{\theta}, \hat{\eta})^2\right] / \mathbb{E}_n(D_i)^2$ *is a consistent estimator of the asymptotic variance.*

A proof of this theorem is given in the appendix. Establishing asymptotic normality of two-step estimators that rely on a regularized first step such as $\hat{\theta}$ has long been a conundrum in this literature. However, Belloni et al. (2012, 2014a, 2017) among other papers by the same authors broke the path to valid post-selection inference by moving from perfect selection to combining estimation of the high-dimensional nuisance parameters with sufficient quality with immunization of the estimating moment for the parameter of interest. Chernozhukov et al. (2015) nicely exposes the theory behind this approach and serves as the main methodological tool behind our proof. Chernozhukov et al. (2018a) extended this approach to machine learning tools in general and further simplified the proofs by proposing sample-splitting. While this more recent contribution is very appealing, it was not available at the time the present paper was conceived.

This theorem relies on estimators of the nuisance parameters that are "good enough" in terms rates of convergence. Theorem 2.3 in the appendix states these rates.

4. Simulations

The aim of this experiment is two-fold: illustrate the better properties of the immunized estimator over the naive plug-in and compare it with other competitors. In particular, we compare it with an inverse propensity score weighting estimator where the propensity score is estimated using a Logit-Lasso Van de Geer (2008) and with a similar estimator proposed by Farrell (2015).

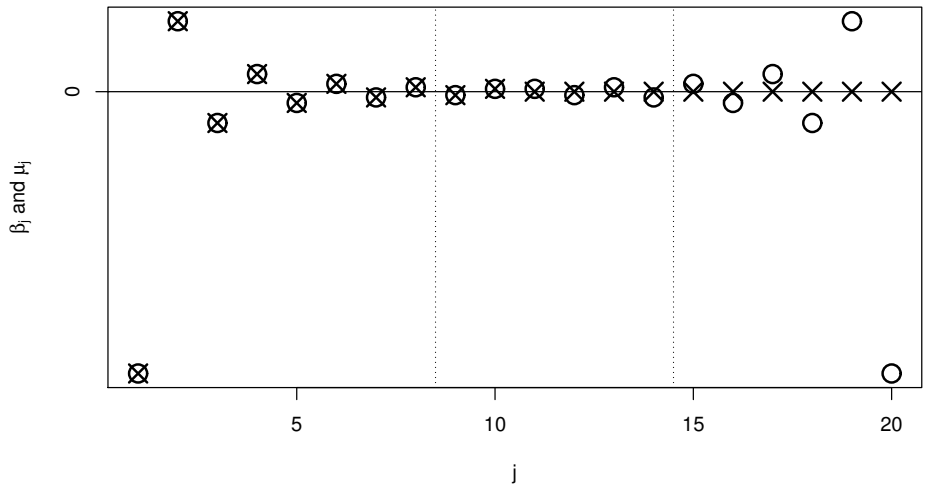
In our main specification, DGP1, the outcome equation is linear and given by: $y_i = d_i\alpha + x_i'\mu + \varepsilon_i$, where $\alpha = 0$, $\varepsilon_i \perp\!\!\!\perp x_i$, and $\varepsilon_i \sim \mathcal{N}(0,1)$. The treatment equation follows a Probit model, $d_i \sim \text{Probit}\,(x_i'\gamma)$. The covariates are simulated as $x_i \sim \mathcal{N}(0,\Sigma)$, where each entry of the variance-covariance matrix is set as follows: $\Sigma_{j,k} = .5^{|j-k|}$. The most

interesting part of the DGP is the form of the coefficients γ and μ:

$$\gamma_j = \begin{cases} \rho_d(-1)^j/j^2, \, j < p/2 \\ 0, \text{ elsewhere} \end{cases}, \, \mu_j = \begin{cases} \rho_y(-1)^j/j^2, \, j < p/2 \\ \rho_y(-1)^{j+1}/(p-j+1)^2, \text{ elsewhere} \end{cases}$$

We are in an approximately sparse setting for both equations. ρ_y and ρ_d are constants that fix the signal-to-noise ratio, in the sense that a larger ρ_y means that covariates play a larger role. The trick here is that some variables that matter a lot in the outcome equation are irrelevant in the treatment assignment rule. The fact that the balancing program will miss some relevant variables for the outcome should create a bias, a non-Normal behavior or at least a wider variance. Figure 2.1 depicts the sparsity pattern of both coefficients for $p = 20$.

Figure 2.1: Sparsity patterns of β (crosses) and μ (circles)



Note: In this example, $\rho_d = \rho_y = 1$. The central region of the graph represents the coefficients γ and μ associated with variables that do not play an important role in either the equation equation or the outcome equation. The left region shows the coefficients associated with variables that are important for both equations. In the right region, only μ is different from 0, meaning that the variables determine the outcome equation but not the selection equation.

We expect a naive plug-in estimator to miss the variables located at the far-right of the plot, thereby creating a bias in the treatment effect estimate. The immunized procedure is expected to correct for this bias.

We explore three alternatives to DGP1. DGP2 is similar in all respects, except that μ_j is equal to 0 for all j. In other words, the outcome equation does not depend on the covariates. In this specification, the naive plug-in estimator is theoretically correct and we can check whether the immunized estimator performs as well as the naive version. DGP3 explores a situation with a heterogeneous treatment effect: the outcome equation

is specified as $y_i = d_i\alpha + x_i'\mu + d_i x_i'\gamma + \varepsilon_i$, with $\gamma_j = 10$ for all j. Since $x_i \sim \mathcal{N}(0, \Sigma)$, this setting still yields an ATT equal to zero. DGP4 relaxes the linearity of the outcome equation and allows to check for double robustness. In this DGP, we specify the outcome equation as follows: $y_i = d_i\alpha + (x_i'\mu)^2 + \varepsilon_i$, but only the covariates x_i are used in the estimation procedure.

Tables 2.1 to 2.4 display the results of Monte-Carlo simulations for both the naive plug-in estimator and the immunized estimator, compared with 3 potential alternatives: inverse propensity score weighting and Farrell (2015) estimator (with either Lasso or post-Lasso procedures). We present results of 1,000 replications with each DGP introduced in the previous paragraph, for values of n and p varying between 50 and 500. All other parameters are held fixed.

The first striking characteristic of these results is that in our baseline DGP (Table 2.1), the bias of the plug-in estimator is almost always larger in absolute value than the bias of the immunized estimator, as predicted. In addition, the difference in bias grows with the sample size. For example, for $n = 500$, the bias of the plug-in estimator is about twice as big as the bias of the immunized estimator. Similarly, the root mean squared error (RMSE) is always higher for the naive estimator than for the immunized one. The difference again increases with sample size. The p-value of the Shapiro test is also usually quite high, showing that the null hypothesis of normality cannot be rejected. These results illustrate the theoretical asymptotic properties of our estimator. Moreover, Table 2.2 also shows that in a setting in which the naive plug-in would be appropriate, the immunized estimator performs as well as the naive estimator. Table 2.3 shows that these findings are robust to heterogeneous treatment effects, as both estimators perform exactly the same way with or without heterogeneity.

How does the immunized estimator fare compared with alternative methods? Tables 2.1 to 2.4 also show the performance of four alternative methods. The first one is inverse propensity-score weighting, estimating the propensity score with a Logit-Lasso. Overall, this method performs similarly to the naive plug-in estimator. This is not surprising, as it is likely to suffer from the same bias as the latter. The second alternative is the method introduced by Belloni et al. (2014a) (denoted as "BCH"). The main difference between this estimator and the immunized estimator is that the former relies on the assumption that the treatment effect is homogeneous. We thus expect the third DGP to yield better results for the immunized estimator than for BCH. Indeed, Table 2.3 shows that BCH has both a very high bias and RMSE in this case. As an order of magnitude, they tend to be 10 times as high as for the immunized estimator, whatever the size of the sample or the number of covariates. In addition, the null hypothesis of the Shapiro test tends to be rejected most of the times.

The third and fourth alternatives are the Farrell (2015) estimator in two versions: one based on a Lasso procedure (denoted as "Farrell" in our tables), and one based on a post-Lasso procedure ("Farrell PL"). The theory suggests that the latter tends to have a smaller bias, because it removes undue shrinkage introduced by the ℓ_1-regularization in the first step. For DGP1 to DGP3, we find that the immunized estimator performs similarly to Farrell's method, and that the post-Lasso procedure indeed reduces the bias

Table 2.1: Monte-Carlo Simulations (DGP1)

	$p = 50$			$p = 100$			$p = 200$			$p = 500$		
	RMSE	Bias	Shapiro	RMSE	Bias	Shapiro	RMSE	Bias	Shapiro	RMSE	Bias	Shapiro
$n = 50$												
Naive Plug-in	0.993	0.829	0.326	1.060	0.891	0.831	–	–	–	–	–	–
immunized	0.689	0.474	0.010	0.793	0.589	0.016	–	–	–	–	–	–
Inv. prop. weighting	0.925	0.771	0.491	0.99	0.828	0.28	–	–	–	–	–	–
BCH	0.873	0.567	0.074	1.048	0.811	0.715	–	–	–	–	–	–
Farrell	0.710	0.509	0.006	0.816	0.616	0.008	–	–	–	–	–	–
Farrell PL	0.657	0.210	0.000	0.794	0.301	0.000	–	–	–	–	–	–
$n = 100$												
Naive Plug-in	0.816	0.719	0.649	0.824	0.739	0.462	0.843	0.758	0.710	–	–	–
immunized	0.396	0.264	0.002	0.395	0.270	0.030	0.424	0.300	0.002	–	–	–
Inv. prop. weighting	0.749	0.653	0.900	0.761	0.674	0.488	0.780	0.697	0.591	–	–	–
BCH	0.248	0.057	0.318	0.253	0.042	0.087	0.299	0.076	0.000	–	–	–
Farrell	0.396	0.266	0.008	0.406	0.282	0.011	0.440	0.318	0.000	–	–	–
Farrell PL	0.323	0.102	0.000	0.339	0.104	0.000	0.414	0.144	0.000	–	–	–
$n = 200$												
Naive Plug-in	0.616	0.556	0.659	0.637	0.581	0.844	0.645	0.591	0.853	0.678	0.628	0.383
immunized	0.249	0.160	0.013	0.261	0.174	0.360	0.264	0.179	0.005	0.277	0.204	0.958
Inv. prop. weighting	0.571	0.507	0.984	0.587	0.529	0.98	0.596	0.539	0.849	0.627	0.576	0.575
BCH	0.176	0.055	0.010	0.177	0.056	0.725	0.176	0.056	0.536	0.178	0.061	0.311
Farrell	0.241	0.154	0.161	0.253	0.168	0.743	0.262	0.179	0.022	0.277	0.206	0.798
Farrell PL	0.206	0.062	0.000	0.233	0.076	0.198	0.251	0.103	0.019	0.297	0.140	0.000
$n = 500$												
Naive Plug-in	0.434	0.399	0.712	0.431	0.398	0.560	0.452	0.421	0.430	0.469	0.437	0.576
immunized	0.154	0.099	0.825	0.151	0.102	0.975	0.159	0.110	0.330	0.162	0.116	0.819
Inv. prop. weighting	0.397	0.359	0.803	0.397	0.360	0.765	0.418	0.383	0.802	0.436	0.402	0.275
BCH	0.120	0.051	0.514	0.115	0.056	0.752	0.118	0.054	0.617	0.115	0.055	0.135
Farrell	0.144	0.088	0.791	0.143	0.092	0.980	0.150	0.100	0.372	0.155	0.108	0.910
Farrell PL	0.121	0.026	0.803	0.121	0.034	0.403	0.130	0.042	0.137	0.142	0.059	0.478

Note: Results based on 1,000 replications. The data generating process is such that the R^2 of the outcome equation is .8 and the R^2 of the treatment equation is .2. The naive plug-in and immunized estimates are computed with parameters $c_y = .2$ and $c_d = .7$. The inverse propensity score weighting estimates are computed by estimating the propensity score using a Logit-Lasso algorithm. The Shapiro columns display the p-value of the Shapiro test (the null hypothesis is normality of the distribution).

Table 2.2: Monte-Carlo Simulations (DGP2: Outcome Independent from X)

	$p = 50$			$p = 100$			$p = 200$			$p = 500$		
	RMSE	Bias	Shapiro	RMSE	Bias	Shapiro	RMSE	Bias	Shapiro	RMSE	Bias	Shapiro
$n = 50$												
Naive Plug-in	0.287	0.004	0.484	0.293	0.001	0.102	–	–	–	–	–	–
immunized	0.295	0.002	0.572	0.300	0.002	0.187	–	–	–	–	–	–
Inv. prop. weighting	0.273	0.005	0.424	0.279	0.001	0.037	–	–	–	–	–	–
BCH	0.293	0.000	0.347	0.293	0.002	0.115	–	–	–	–	–	–
Farrell	0.293	0.001	0.562	0.298	0.002	0.191	–	–	–	–	–	–
Farrell PL	0.446	-0.012	0.000	0.475	-0.009	0.000	–	–	–	–	–	–
$n = 100$												
Naive Plug-in	0.204	-0.001	0.364	0.208	-0.009	0.427	0.206	-0.002	0.476	–	–	–
immunized	0.206	0.000	0.339	0.211	-0.009	0.208	0.209	-0.001	0.306	–	–	–
Inv. prop. weighting	0.195	-0.001	0.208	0.199	-0.009	0.481	0.197	-0.003	0.531	–	–	–
BCH	0.215	0.000	0.189	0.215	-0.010	0.586	0.211	-0.002	0.826	–	–	–
Farrell	0.204	-0.001	0.436	0.210	-0.010	0.270	0.209	-0.002	0.383	–	–	–
Farrell PL	0.266	-0.003	0.000	0.290	-0.011	0.000	0.386	-0.022	0.000	–	–	–
$n = 200$												
Naive Plug-in	0.146	0.000	0.071	0.150	0.000	0.552	0.148	0.005	0.712	0.147	0.007	0.945
immunized	0.147	-0.001	0.129	0.152	0.000	0.442	0.149	0.005	0.679	0.148	0.008	0.955
Inv. prop. weighting	0.141	0.000	0.217	0.146	0.000	0.57	0.141	0.005	0.684	0.140	0.007	0.726
BCH	0.152	0.000	0.077	0.151	0.000	0.901	0.151	0.004	0.346	0.153	0.006	0.345
Farrell	0.147	0.000	0.211	0.152	0.000	0.444	0.148	0.005	0.673	0.148	0.008	0.794
Farrell PL	0.180	0.002	0.000	0.196	0.001	0.172	0.203	0.005	0.320	0.224	0.002	0.000
$n = 500$												
Naive Plug-in	0.098	0.001	0.598	0.092	-0.002	0.468	0.095	-0.002	0.624	0.093	-0.001	0.976
immunized	0.098	0.001	0.572	0.092	-0.001	0.319	0.095	-0.001	0.618	0.093	-0.001	0.979
Inv. prop. weighting	0.095	0.001	0.552	0.089	-0.002	0.499	0.092	-0.001	0.784	0.090	-0.001	0.960
BCH	0.099	0.001	0.732	0.092	0.001	0.543	0.095	-0.001	0.157	0.094	-0.001	0.280
Farrell	0.098	0.001	0.552	0.092	-0.002	0.380	0.096	-0.001	0.630	0.093	-0.001	0.981
Farrell PL	0.110	0.000	0.822	0.106	-0.002	0.519	0.116	-0.001	0.555	0.119	-0.001	0.546

Note: Results based on 1,000 replications. The data generating process is such that the R^2 of the outcome equation is .8 and the R^2 of the treatment equation is .2. The naive plug-in and immunized estimates are computed with parameters $c_y = .2$ and $c_d = .7$. The inverse propensity score weighting estimates are computed by estimating the propensity score using a Logit-Lasso algorithm. The Shapiro columns display the p-value of the Shapiro test (the null hypothesis is normality of the distribution).

Table 2.3: Monte-Carlo Simulations (DGP3: Heterogeneous Treatment Effect)

	$p = 50$			$p = 100$			$p = 200$			$p = 500$		
	RMSE	Bias	Shapiro	RMSE	Bias	Shapiro	RMSE	Bias	Shapiro	RMSE	Bias	Shapiro
$n = 50$												
Naive Plug-in	0.993	0.829	0.326	1.060	0.891	0.831	–	–	–	–	–	–
immunized	0.689	0.474	0.010	0.793	0.589	0.016	–	–	–	–	–	–
Inv. prop. weighting	0.925	0.771	0.491	0.990	0.828	0.280	–	–	–	–	–	–
BCH	7.772	2.303	0.000	7.709	1.449	0.000	–	–	–	–	–	–
Farrell	0.710	0.509	0.006	0.816	0.616	0.008	–	–	–	–	–	–
Farrell PL	0.657	0.210	0.000	0.794	0.301	0.000	–	–	–	–	–	–
$n = 100$												
Naive Plug-in	0.816	0.719	0.649	0.824	0.739	0.462	0.843	0.758	0.710	–	–	–
immunized	0.396	0.264	0.002	0.395	0.270	0.030	0.424	0.300	0.002	–	–	–
Inv. prop. weighting	0.749	0.653	0.900	0.761	0.674	0.488	0.780	0.697	0.591	–	–	–
BCH	9.512	5.468	0.000	10.32	4.331	0.000	11.979	3.66	0.000	–	–	–
Farrell	0.396	0.266	0.008	0.406	0.282	0.011	0.440	0.318	0.000	–	–	–
Farrell PL	0.323	0.102	0.000	0.339	0.104	0.000	0.414	0.144	0.000	–	–	–
$n = 200$												
Naive Plug-in	0.616	0.556	0.659	0.637	0.581	0.844	0.645	0.591	0.853	0.678	0.628	0.383
immunized	0.249	0.160	0.013	0.261	0.174	0.360	0.264	0.179	0.005	0.277	0.204	0.958
Inv. prop. weighting	0.571	0.507	0.984	0.587	0.529	0.98	0.596	0.539	0.849	0.627	0.576	0.575
BCH	9.275	6.688	0.000	10.049	6.676	0.000	12.283	7.141	0.000	16.616	7.341	0.000
Farrell	0.241	0.154	0.161	0.253	0.168	0.743	0.262	0.179	0.022	0.277	0.206	0.798
Farrell PL	0.206	0.062	0.000	0.233	0.076	0.198	0.251	0.103	0.019	0.297	0.140	0.000
$n = 500$												
Naive Plug-in	0.434	0.399	0.712	0.431	0.398	0.560	0.452	0.421	0.430	0.469	0.437	0.576
immunized	0.154	0.099	0.825	0.151	0.102	0.975	0.159	0.110	0.330	0.162	0.116	0.819
Inv. prop. weighting	0.397	0.359	0.803	0.397	0.360	0.765	0.418	0.383	0.802	0.436	0.402	0.275
BCH	8.163	5.970	0.171	8.729	6.148	0.647	9.664	7.291	0.379	11.467	7.182	0.045
Farrell	0.144	0.088	0.791	0.143	0.092	0.980	0.150	0.100	0.372	0.155	0.108	0.910
Farrell PL	0.121	0.026	0.803	0.121	0.034	0.403	0.13	0.042	0.137	0.142	0.059	0.478

Note: Results based on 1,000 replications. The data generating process is such that the R^2 of the outcome equation is .8 and the R^2 of the treatment equation is .2. The naive plug-in and immunized estimates are computed with parameters $c_y = .2$ and $c_d = .7$. The inverse propensity score weighting estimates are computed by estimating the propensity score using a Logit-Lasso algorithm. The Shapiro columns display the p-value of the Shapiro test (the null hypothesis is normality of the distribution).

Table 2.4: Monte-Carlo simulations (DGP4: Non-Linear Outcome Equation)

	$p = 50$			$p = 100$			$p = 200$			$p = 500$		
	RMSE	Bias	Shapiro	RMSE	Bias	Shapiro	RMSE	Bias	Shapiro	RMSE	Bias	Shapiro
$n = 50$												
Naive Plug-in	1.692	0.119	0.749	1.676	0.200	0.088	–	–	–	–	–	–
immunized	1.720	0.052	0.375	1.710	0.094	0.026	–	–	–	–	–	–
Inv. prop. weighting	1.734	0.487	0.864	1.739	0.590	0.057	–	–	–	–	–	–
BCH	1.669	-0.052	0.853	1.657	0.073	0.030	–	–	–	–	–	–
Farrell	1.703	0.021	0.345	1.689	0.084	0.073	–	–	–	–	–	–
Farrell PL	2.812	-0.164	0.000	3.346	-0.112	0.000	–	–	–	–	–	–
$n = 100$												
Naive Plug-in	1.148	0.218	0.947	1.152	0.263	0.004	1.195	0.224	0.420	–	–	–
immunized	1.158	0.209	0.843	1.168	0.245	0.011	1.207	0.172	0.200	–	–	–
Inv. prop. weighting	1.256	0.548	0.499	1.271	0.607	0.012	1.315	0.602	0.179	–	–	–
BCH	1.089	-0.037	0.338	1.089	0.032	0.009	1.130	0.032	0.052	–	–	–
Farrell	1.163	0.176	0.946	1.142	0.205	0.027	1.199	0.130	0.084	–	–	–
Farrell PL	1.911	0.045	0.000	1.852	0.037	0.000	2.482	-0.165	0.000	–	–	–
$n = 200$												
Naive Plug-in	0.935	0.328	0.694	0.847	0.280	0.091	0.863	0.289	0.399	0.889	0.249	0.008
immunized	0.952	0.332	0.318	0.862	0.280	0.376	0.869	0.279	0.380	0.895	0.210	0.025
Inv. prop. weighting	1.075	0.597	0.652	0.988	0.576	0.047	1.025	0.614	0.279	1.058	0.629	0.009
BCH	0.799	0.008	0.723	0.772	-0.005	0.162	0.760	-0.014	0.959	0.810	-0.020	0.224
Farrell	0.971	0.322	0.919	0.876	0.265	0.188	0.870	0.252	0.132	0.888	0.180	0.045
Farrell PL	1.438	0.171	0.000	1.403	0.097	0.000	1.485	0.06	0.000	1.36	0.031	0.000
$n = 500$												
Naive Plug-in	0.634	0.286	0.191	0.625	0.282	0.007	0.618	0.276	0.204	0.616	0.296	0.299
immunized	0.637	0.288	0.188	0.626	0.280	0.007	0.621	0.274	0.211	0.613	0.288	0.338
Inv. prop. weighting	0.768	0.487	0.001	0.764	0.502	0.000	0.770	0.514	0.067	0.791	0.561	0.070
BCH	0.504	-0.002	0.108	0.498	0.013	0.801	0.491	-0.019	0.129	0.479	0.006	0.324
Farrell	0.722	0.337	0.043	0.689	0.315	0.005	0.682	0.300	0.092	0.658	0.302	0.145
Farrell PL	0.882	0.165	0.000	0.829	0.136	0.000	0.890	0.133	0.000	0.867	0.119	0.000

Note: Results based on 1,000 replications. The data generating process is such that the R^2 of the outcome equation is .8 and the R^2 of the treatment equation is .2. The naive plug-in and immunized estimates are computed with parameters $c_y = .2$ and $c_d = .7$. The inverse propensity score weighting estimates are computed by estimating the propensity score using a Logit-Lasso algorithm. The Shapiro columns display the p-value of the Shapiro test (the null hypothesis is normality of the distribution).

compared to our estimator. However, it has two drawbacks. First, the Shapiro test tends to be rejected more often for the post-Lasso procedure than for the immunized estimator. Second, Table 2.4 shows that Farrell's methods are not as robust to non-linear outcome equations as the immunized. Indeed, with DGP4 we find that the RMSE is systematically smaller for the immunized than for Farrell's estimators. In particular, the post-Lasso procedure yields much higher RMSE for small samples (twice as high for $n = 50$). This illustrates the advantages of the immunized estimator double robustness.

5. Empirical Applications

5.1. Job Training Program, LaLonde (1986)

We revisit LaLonde (1986). This dataset was first built to assess the impact of the National Supported Work (NSW) program. The NSW is a transitional, subsidized work experience program targeted towards people with longstanding employment problems: ex-offenders, former drug addicts, women who were long-term recipients of welfare benefits and school dropouts. Here, the quantity of interest is the ATET, defined as the impact of the participation in the program on 1978 yearly earnings in dollars. The treated group gathers people who were randomly assigned to this program from the population at risk ($n_1 = 185$).Two control groups are available. The first one is experimental: it is directly comparable to the treated group as it has been generated by a random control trial (sample size $n_0 = 260$). The second one comes from the Panel Study of Income Dynamics (PSID) (sample size $n_0 = 2490$). The presence of the experimental sample allows to obtain a benchmark for ATET obtained with observational data. We use these datasets to compare our estimator with other competitors and defer discussion of the NSW program and the controversy regarding econometric estimates of nonexperimental causal effects to the paper by LaLonde (1986) and subsequent contributions by Dehejia and Wahba (2002); Smith and Todd (2005).

To allow for a flexible specification, we consider the setting of Farrell (2015) and take the raw covariates of the dataset (age, education, black, hispanic, married, no degree, income in 1974, income in 1975, no earnings in 1974, no earnings in 1975), two-by-two-interactions between the four continuous variables and the dummies, two-by-two interactions between the dummies and up to a degree of order 5 polynomial transformations of continuous variables. Continuous variables are linearly rescaled to $[0, 1]$. All in all, we end up with 172 variables to select from. The experimental benchmark for the ATT estimate is \$1,794 (671). We compare several estimators: the naive plug-in estimator, the immunized plug-in estimator, the doubly-robust estimator of Farrell (2015), the double-post-selection linear estimator of Belloni et al. (2014b), and a simple OLS estimator where all the covariates are included.

Table 2.5 displays the results. Columns (3)-(5) show estimators that give a credible value for the ATT with respect to the experimental benchmark. However, they differ in their variances as one can easily see. Farrell (2015) in its Lasso version and the immunized estimator achieve the lowest standard-error. Notably, Farrell (2015) in its Lasso version and the immunized estimator are the only ones out of six estimators which

display a significant, positive impact similarly to the experimental benchmark. The immunized estimator estimator offers a large improvement on bias and standard error over the naive plug-in estimator, which augments the evidence given by the Monte Carlo experiment. The estimate obtained using Farrell (2015) shown in the Table differ from the on displayed in the original paper because we have not automatically included the variables *education, 1974 income* and *nodegree* in the set of theory pre-selected covariates as it is done in the original paper. When doing so, the results are slightly better but not qualitatively different for this estimator, but we thought it would bias the comparison as other estimators do not include a set of pre-selected variables. For estimators from columns (2) to (6), the penalty parameters can potentially be tuned to obtain a better bias-variance trade-off. The OLS estimator in column (7) presents a benchmark of a very simple model that does not use any selection at all.

Table 2.5: Average Treatment Effect on the Treated for NSW.

	Estimator:			
	Experimental benchmark (1)	*Plug-In Naive* (2)	*Immunized Estimator* (3)	*Farrell (2015) Lasso* (4)
Estimate	1,794.34	214.72	1,495.91	1,537.80
Standard error (Asy.)	(671.00)	(873.88)	(705.32)	(675.16)
.95 confidence interval (Asy)	[519;3,046]	[-1,498;1,928]	[114;2,878]	[214;2,861]
# variables in Propensity Score	none	8	8	3
# variables in Outcome function	none	none	11	16
	Farrell (2015) Post-Lasso (5)	*BCH (2014)* - (6)	*OLS* - (7)	
Estimate	1,340.24	382.28	83.17	
Standard error (Asy.)	(778.38)	(852.40)	(1,184.48)	
.95 confidence interval (Asy)	[-185;2,866]	[-1,288;2,053]	[-2,238;2,405]	
# variables in propensity score	3	8	none	
# variables in regression function	16	10	172	

Note: The experimental estimate is computed on experimental data, column (1). (Asy.) signals the asymptotic approximation estimator of the quantity is used.

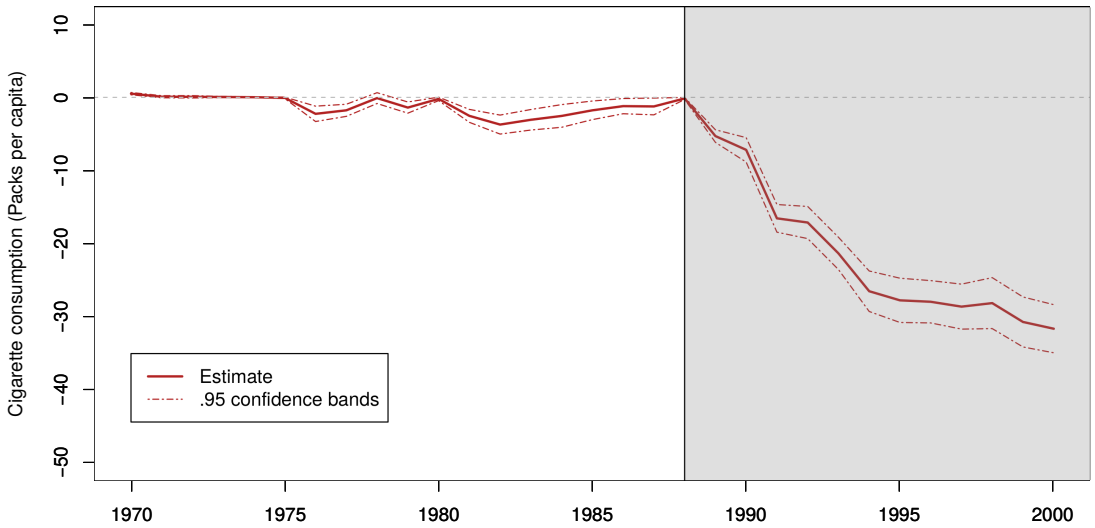
5.2. California Tobacco Control Program, Abadie et al. (2010)

Proposition 99 is one of the first and most ambitious large-scale tobacco control program, implemented in 1989 in California. It includes a vast array of measures, including an increase in cigarette taxation of 25 cents per pack, and a significant effort in prevention and education. In particular, the tax revenues generated by Proposition 99 were used to fund anti-smoking campaigns. Abadie et al. (2010) analyze the impact of the law on tobacco consumption in California. Since this program was only enforced in California, it is a classic example where the synthetic control method applies, and more standard public policy evaluation tools cannot be used. It is possible to reproduce a synthetic

California by reweighting other states so as to imitate California's behavior.

For this purpose, Abadie et al. (2010) consider the following covariates: retail price of cigarettes, state log income per capita, percentage of population between 15-24, per capita beer consumption (all 1980-1988 averages). 1970 to 1975, 1980 and 1988 cigarette consumptions are also included. Using the same variables, we conduct the same analysis with our estimator. Figure 2.2 displays the estimated effect of Proposition 99 using the immunized estimator.

Figure 2.2: The effect of Proposition 99 on per capita tobacco consumption.



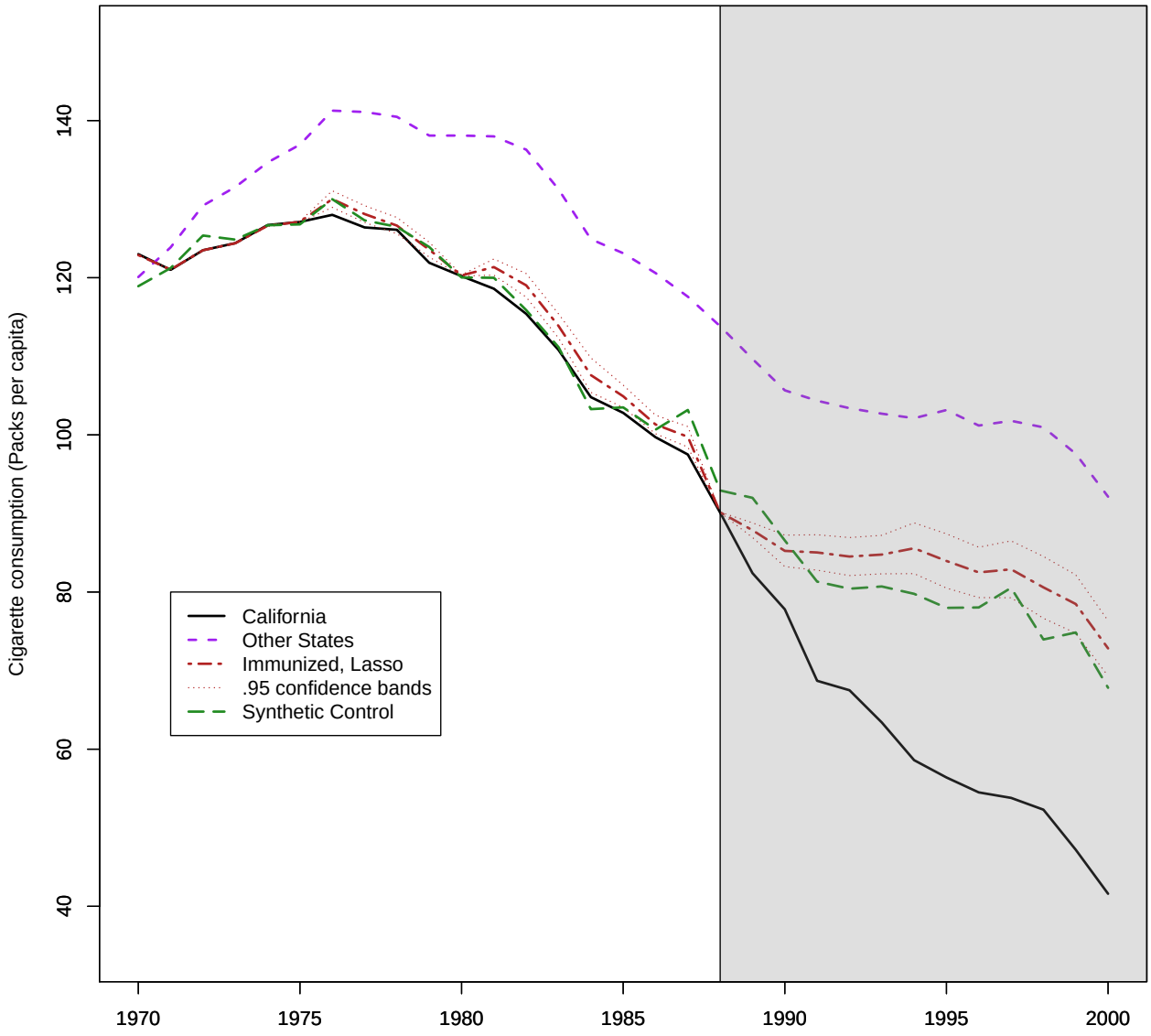
Note: The shaded area represents the post-treatment period. The .95 confidence interval is based on the asymptotic approximation.

We finds almost no effect of the policy over the pre-treatment period, giving credibility to the counterfactual employed. A steady decline takes place after 1988, and in the long-run, tobacco consumption is estimated to have decreased by about 30 packs per capita per year in California as a consequence of the policy. The variance is larger towards the end of the period, because covariates are measured in the pre-treatment period and become less relevant as predictors. It is also to be noted that by construction, including 1970 to 1975, 1980 and 1988 cigarette consumptions among the covariates yields an almost perfect fit at theses dates because of the immunization procedure (up to the amount of shrinkage induced by the Lasso).

Finally, Figure 2.3 allows a comparison between the immunized estimator and the synthetic control method. The dashed green line is the synthetic control counterfactual. Notice that they do not exactly match the plots of Abadie et al. (2010), in which the weights given to each predictors are optimized to best fit the outcome over the whole pre-treatment period. Instead, the green curve optimizes the predictor weights only using dates 1970 through 1975, 1980 and 1988. This strategy allows a fairer comparison with our estimator that does not use California's per capita tobacco consumption outside those dates to optimize the fit, while period 1975-1988 can be thought of as a semi-placebo test.

Both our estimator and the synthetic control are credible counterfactuals, as both are

Figure 2.3: Cigarette consumption in California, actual and counterfactual.



Note: The solid black line is California tobacco consumption as in the data. The dotted purple line is a simple average of other U.S. states. The dashed red line is the immunized estimator as presented in the paper along with the .95 confidence bands. The dashed green line is synthetic California.

able to closely match California pre-treatment tobacco consumption. They both offer a sizable improvement over a sample average over the U.S. that did not implement any tobacco control program. Furthermore, even if our estimator gives a result relatively similar to the synthetic control, it displays a smoother pattern especially towards the end of the 1980s. The estimated treatment effect appears to be larger with the immunized estimate than with the synthetic control. However, it is hard to conclude that this difference is significant because one cannot easily compute confidence intervals for the synthetic control estimates, without making stringent assumptions about the program effect. The availability of standard asymptotic approximation for confidence intervals is to the advantage of our method.

6. Conclusion

This chapter proposed a parametric generalization of the synthetic control method, which is developed both in the usual asymptotic framework and in the conventional high-dimensional one. The basic idea to deal with the high-dimensionality inherent to the synthetic control method is to move from choosing a weight for each non-treated unit to choosing a weight for each covariates, by linking individual weights to covariates through the propensity score.

The proposed estimator is doubly robust, consistent and asymptotically normal uniformly over a large class of data-generating processes. It is also immunized against first-step selection mistakes. We illustrated these properties using Monte Carlo simulations and applications to both standard and potentially high-dimensional settings, and offer a comparison with the synthetic control method.

7. Appendix A: Algorithm for Feasible Penalty Loadings

The ideal penalty loadings for estimation of β_0 are given by:

$$\lambda^d := c\Phi^{-1}(1-\gamma/2p)/\sqrt{n}$$
$$\psi_j^d := \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(1-D_i)h(X_i'\beta_0)-D_i\right]^2 X_{i,j}^2}, \text{ for } j=1,...,p$$

The ideal penalty loadings for estimation of μ_0 are given by:

$$\lambda^d := c\Phi^{-1}(1-\gamma/2p)/\sqrt{n}$$
$$\psi_j^y := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(1-D_i)h'(X_i^T\beta_0)^2\left[Y_i-X_i^T\mu_0\right]^2 X_{i,j}^2}, \text{ for } j=1,...,p$$

where $c>1$ is an absolute constant, $\gamma \lesssim \log(p\vee n)$ and β_0 and μ_0 are the true coefficients. We follow Belloni et al. (2012) a,d set $\gamma := .1/\log(p\vee n)$ and $c := 1.1$.

For estimating the penalty loadings $\left\{\psi_j^d\right\}_{j=1}^d$ in the calibration part, we use the following algorithm:

Set a small constant $v>0$ and a maximal number of iterations K.

1. Start by using a preliminary estimate $\beta^{(0)}$ of β_0. For example set $\beta^{(0)}$ with its first entries equal to $\log(n_1/n_0)$ and all other entries equal to zero. Then set $\tilde{\psi}_j^{(0)} = \sqrt{\mathbb{E}_n\left[(1-D_i)h(X_i^T\beta^{(0)})-D_i\right]^2 X_{i,j}^2}$, $j=1,...,p$.
   At step k, set $\tilde{\psi}_j^{(k)} = \sqrt{\mathbb{E}_n\left[(1-D_i)h(X_i^T\beta^{(k)})-D_i\right]^2 X_{i,j}^2}$, $j=1,...,p$.
2. Estimate the model by the Calibration Lasso of equation 2.7 using the overall penalty level λ and penalty loadings found previously, to obtain $\hat{\beta}^{(k)}$.
3. Stop if $\max_{j=1,...,p}|\tilde{\psi}_j^{(k)}-\tilde{\psi}_j^{(k-1)}| \leq v$ or $k>K$. Set k=k+1 and go to step 1 otherwise.

Asymptotic validity of this approach is established in (Belloni et al., 2012, Lemma 11). The penalty loadings estimation of the immunization step follows a similar procedure. In this specific case, replace β_0 by $\hat{\beta}$ obtained in the calibration step.

8. Appendix B: Proofs

Proof of Theorem 2.1: First, note that β_0 satisfies $\mathbb{E}\left[(1-D)h(X'\beta_0)X\right] = \mathbb{E}\left[DX\right]$. As a result, for any $\mu\in\mathbb{R}^p$,

$$\mathbb{E}\left[\left(D-(1-D)h(X'\beta_0)\right)\left(Y-X'\mu\right)\right] = \mathbb{E}\left[\left(D-(1-D)h(X'\beta_0)\right)Y\right].$$

Because $(1-D)Y = (1-D)Y_0$ and $DY = DY_1$, θ_0 verifies the moment condition 2.5 if and only if:

$$\mathbb{E}\left[h(X'\beta_0)Y(1-D)\right] = \mathbb{E}\left[DY_0\right].$$

The mean independence assumption allows to write:

$$\mathbb{E}\left[h(X'\beta_0)Y(1-D)\right] = \mathbb{E}\left[h(X'\beta_0)\mathbb{E}(1-D|X)\mathbb{E}(Y_0|X)\right].$$

We consider the two cases separately.

1. The linear case: $\mathbb{E}(Y_0|X) = X'\mu_0$.

$$\begin{aligned}\mathbb{E}\left[h(X'\beta_0)Y(1-D)\right] &= \mathbb{E}\left[h(X'\beta_0)(1-D)X'\mu_0\right] \\ &= \mathbb{E}\left[DX'\mu_0\right] \\ &= \mathbb{E}\left[D\mathbb{E}(Y_0|X)\right] \\ &= \mathbb{E}\left[DY_0\right].\end{aligned}$$

   The first equality uses the Mean Independence assumption. The second line follows from the fact that β_0 is such that $\mathbb{E}[(1-D)h(X'\beta_0)X] = \mathbb{E}[DX]$.

2. Propensity score given by $P(D=1|X) = G(X'\beta_0)$.

$$\begin{aligned}\mathbb{E}\left[h(X'\beta_0)Y(1-D)\right] &= \mathbb{E}\left[h(X'\beta_0)(1-G(X'\beta_0))\mathbb{E}(Y_0|X)\right] \\ &= \mathbb{E}\left[G(X'\beta_0)\mathbb{E}(Y_0|X)\right] \\ &= \mathbb{E}\left[\mathbb{E}(D|X)\mathbb{E}(Y_0|X)\right] \\ &= \mathbb{E}\left[DY_0\right].\end{aligned}$$

□

Proof of Theorem 2.2: Denote the observed data by $Z_i := (Y_i, D_i, X_i)$ and let $\eta := (\beta', \mu')'$ denote the parameter gathering the two nuisance vectors. The estimating moment for θ_0 is $g(Z, \theta, \eta) := [D - (1-D)h(X'\beta)][Y - X'\mu] - D\theta$ and π_0 the probability of being treated: $\pi_0 := \mathbb{P}(D=1)$. Recall that we define (θ_0, η_0) as the values satisfying:

$$\mathbb{E}g(Z, \theta_0, \eta_0) = 0.$$

All these quantities implicitly depends on the sample size n, but we suppress it in the notations when obvious.

By linearity of the estimating function g in θ and using Lemma 2.2, there exists $t \in (0,1)$ such that:

$$\begin{aligned}\mathbb{E}_n g(Z, \hat{\theta}, \hat{\eta}) &= \mathbb{E}_n g(Z, \theta_0, \hat{\eta}) + \hat{\pi}(\theta_0 - \hat{\theta}) \\ &= \hat{\pi}(\theta_0 - \hat{\theta}) + \mathbb{E}_n g(Z, \theta_0, \eta_0) + (\hat{\eta} - \eta_0)'\mathbb{E}_n \partial_\eta g(Z, \theta_0, \eta_0)\end{aligned}$$

$$+\frac{1}{2}(\hat{\eta}-\eta_0)'\mathbb{E}_n\partial_\eta\partial_{\eta'}g(Z,\theta_0,\tilde{\eta})(\hat{\eta}-\eta_0),$$

with $\tilde{\eta} := t\eta_0 + (1-t)\hat{\eta}$. The immunized estimator satisfies $\mathbb{E}_n g(Z,\hat{\theta},\hat{\eta}) = 0$. Thus, we obtain

$$\begin{aligned}\hat{\pi}\sqrt{n}(\hat{\theta}-\theta_0) = &\underbrace{\sqrt{n}\mathbb{E}_n g(Z_i,\theta_0,\eta_0)}_{:=I_1} + \underbrace{\sqrt{n}(\hat{\eta}-\eta_0)'\mathbb{E}_n\partial_\eta g(Z_i,\theta_0,\eta_0)}_{:=I_2}\\ &+\underbrace{2^{-1}\sqrt{n}(\hat{\eta}-\eta_0)'\mathbb{E}_n\partial_\eta\partial_{\eta'}g(Z_i,\theta_0,\tilde{\eta})(\hat{\eta}-\eta_0)}_{:=I_3}.\end{aligned}$$

We now prove that I_1 tends to a normal variable (step 1 below), while I_2 and I_3 tend to zero in probability (steps 2 and 3).

Step 1. Let $g_{i,n} := g(Z_i,\theta_0,\eta_0)$ (letting the dependence in n explicit in $g_{i,n}$). Recall that $\mathbb{E}g_{i,n} = 0$. We apply the Lindeberg-Feller Central Limit Theorem for triangular arrays by checking a Lyapunov condition. Because the $(g_{i,n})_{1\leq i\leq n}$ are i.i.d. case, it suffices to prove that

$$\limsup \frac{\mathbb{E}(g_{1,n}^{2+\delta})}{\mathbb{E}(g_{1,n}^2)^{1+\delta/2}} < \infty. \tag{B.1}$$

Using Assumption 2.4, $\mathbb{E}(g_{1,n}^{2+\delta})$ is bounded from above. Moreover,

$$\mathbb{E}(g_{1,n}^2) = \pi_0\mathbb{E}\left((Y_1 - X'\mu_0 - \theta_0)^2|D=1\right) + (1-\pi_0)\mathbb{E}\left(h(X'\beta_0)^2(Y_0-X'\mu_0)^2|D=0\right).$$

Hence, $\liminf \mathbb{E}(g_{1,n}^2) > 0$. Thus, (B.1) holds, and I_1 is asymptotically normal.

Step 2. The derivatives of the estimating moment with respect to the nuisance parameters are

$$\begin{aligned}\frac{\partial}{\partial\beta}g(Z,\theta,\eta) &= -(1-D)h'(X'\beta)\left[Y-X'\mu\right]X,\\ \frac{\partial}{\partial\mu}g(Z,\theta,\eta) &= -\left[D-(1-D)h(X'\beta)\right]X.\end{aligned}$$

Define the random vector U_i of size $2p$ with each element given by:

$$U_{i,j} := \begin{cases} -(1-D_i)h'(X_i'\beta_0)\left[Y_i - X_i'\mu_0\right]X_{i,j} & \text{if } 1\leq j\leq p\\ -\left[D_i-(1-D_i)h(X_i'\beta_0)\right]X_{i,j} & \text{if } p+1\leq j\leq 2p.\end{cases}$$

Recall that Ψ is a square diagonal matrix of dimension $2p$ with elements $\sqrt{n^{-1}\sum_{i=1}^n U_{i,j}^2}$ on its diagonal. Notice that:

$$\|I_2\|_1 \leq \|\Psi(\hat{\eta}-\eta_0)\|_1\|\Psi^{-1}\sqrt{n}\mathbb{E}_n\partial_\eta g(Z_i,\theta_0,\eta_0)\|_\infty.$$

From the orthogonality conditions and Assumption 2.4, $\mathbb{E}U_{i,j} = 0$ and $\mathbb{E}|U_{i,j}|^3 < \infty$ for any i and any j. By Lemma 5 in Belloni et al. (2012) and the fact that $\Phi^{-1}(1-a) \leq$

$\sqrt{-2\log(2a)}$ for all $a \in]0, 1/2[$, we have, with probability approaching one,

$$\|\Psi^{-1}\sqrt{n}\mathbb{E}_n\partial_\eta g(Z_i, \theta_0, \eta_0)\|_\infty \leq \Phi^{-1}(1-\gamma/4p) \leq \sqrt{2\log(2p/\gamma)}.$$

By Theorem 2.3, $\|\Psi(\hat{\eta}-\eta_0)\|_1 \lesssim (s_\beta + s_\mu)\sqrt{\log(p)/n}$. Then, by the growth condition in Assumption 2.3, I_2 converges to 0 in probability.

Step 3. First, we have:

$$\|I_3\|_1 \leq \frac{\sqrt{n}}{2}\|\hat{\eta}-\eta_0\|_2^2 \sup_{a=(a_1,a_2):\|a\|_2>0, \|a_1\|_0\leq\hat{s}_\beta, \|a_2\|_0\leq\hat{s}_\mu} \frac{\sqrt{a'\mathbb{E}_n\partial_\eta\partial_{\eta'}g(Z_i,\theta_0,\tilde{\eta})a}}{\|a\|_2},$$

where in the supremum $a_1 \in \mathbb{R}^p$, $a_2 \in \mathbb{R}^p$ and $a = (a_1, a_2)$, and we use the notations $\hat{s}_\beta := \|\hat{\beta}\|_0$, $\hat{s}_\mu := \|\hat{\mu}\|_0$. From Theorem 2.3, $\|\hat{\eta}-\eta_0\|_2 \lesssim (\sqrt{s_\beta}+\sqrt{s_\mu})\sqrt{\log(p)/n}$. Thus, by Assumption 2.3-(ii), the first term on the right-hand side tends to 0. Hence, it suffices to prove

$$\sup_{\substack{\|a\|_2>0 \\ \|a_1\|_0\leq\hat{s}_\beta \\ \|a_2\|_0\leq\hat{s}_\mu}} \frac{a'\mathbb{E}_n\partial_\eta\partial_{\eta'}g(Z_i,\theta_0,\tilde{\eta})a}{\|a\|_2^2} \lesssim 1. \tag{B.2}$$

First,

$$\begin{aligned}
&\frac{\partial^2}{\partial\beta\partial\beta'}g(Z,\theta,\eta) = -(1-D)h''(X'\beta)\left[Y - X'\mu\right]XX', \\
&\frac{\partial^2}{\partial\mu\partial\beta'}g(Z,\theta,\eta) = \frac{\partial^2}{\partial\beta\partial\mu'}g(Z,\theta,\eta) = (1-D)h'(X'\beta)XX', \\
&\frac{\partial^2}{\partial\mu\partial\mu'}g(Z,\theta,\eta) = 0.
\end{aligned}$$

As a result,

$$\begin{aligned}
a'\mathbb{E}_n\left[\partial_\eta\partial_{\eta'}g(Z_i,\theta_0,\tilde{\eta})\right]a = {} & a_1'\left[\frac{1}{n}\sum_{i=1}^n(1-D_i)h''(X_i'\tilde{\beta})(Y_i - X_i\tilde{\mu})X_iX_i'\right]a_1 \\
& + 2a_2'\left[\frac{1}{n}\sum_{i=1}^n(1-D_i)h'(X_i'\tilde{\beta})X_iX_i'\right]a_1.
\end{aligned} \tag{B.3}$$

Let us consider the first term on the right-hand-side of (B.3). We have

$$\frac{1}{n}\sum_{i=1}^n(1-D_i)h''(X_i'\tilde{\beta})(Y_i - X_i'\tilde{\mu})X_iX_i' = \mathbf{H}_{1,n} + \mathbf{H}_{2,n} + \mathbf{H}_{3,n} + \mathbf{H}_{4,n},$$

where

$$\mathbf{H}_{1,n} = n^{-1}\sum_{i=1}^n(1-D_i)h''(X_i'\beta_0)(Y_i - X_i'\mu_0)X_iX_i',$$

$$\mathbf{H}_{2,n} = n^{-1}\sum_{i=1}^{n}(1-D_i)h''(X_i'\beta_0)X_i'(\mu_0-\tilde{\mu})X_iX_i',$$
$$\mathbf{H}_{3,n} = n^{-1}\sum_{i=1}^{n}(1-D_i)(h''(X_i'\tilde{\beta})-h''(X_i'\beta_0))(Y_i-X_i'\mu_0)X_iX_i',$$
$$\mathbf{H}_{4,n} = n^{-1}\sum_{i=1}^{n}(1-D_i)(h''(X_i'\tilde{\beta})-h''(X_i'\beta_0))X_i'(\mu_0-\tilde{\mu})X_iX_i'.$$

For any $p\times p$ matrix Q, let us define its m-sparse-norm as:

$$\|Q\|_{sp(m)} := \sup_{\substack{\|\delta\|_0\leq m\\ \|\delta\|_2>0}} \frac{\sqrt{\delta' Q\delta}}{\|\delta\|_2}.$$

Start with $\mathbf{H}_{1,n}$. Assumption 2.5, Lemma 1 and Supplemental Appendix C in Belloni and Chernozhukov (2013) imply that

$$\|\mathbf{H}_{1,n}\|_{sp(\hat{s}_\beta)} \lesssim 1.$$

Consider $\mathbf{H}_{2,n}$:

$$\begin{aligned}|a_1'\mathbf{H}_{2,n}a_1| &\leq \left[\sup_{i=1,\ldots,n}|(1-D_i)X_i'(\mu_0-\tilde{\mu})|\right] a_1'\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)h''(X_i'\beta_0)X_iX_i'\right]a_1\\ &\leq K_n(1-t)^2\|\mu_0-\hat{\mu}\|_2^2 a_1'\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)h''(X_i'\beta_0)X_iX_i'\right]a_1.\end{aligned}$$

Using inequalities Assumptions 2.5 and 2.4 and Lemma 1 and Supplemental Appendix C in Belloni and Chernozhukov (2013), we obtain that:

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(1-D_i)h''(X_i'\beta_0)X_iX_i'\right\|_{sp(\hat{s}_\beta)} \lesssim 1,$$

which implies that

$$\|\mathbf{H}_{2,n}\|_{sp(\hat{s}_\beta)} \xrightarrow{p} 0.$$

Next, consider $\mathbf{H}_{3,n}$:

$$\begin{aligned}|a_1'\mathbf{H}_{3,n}a_1| &\leq \left[\sup_{i=1,\ldots,n}\left|(1-D_i)\left(h''(X_i'\tilde{\beta})-h''(X_i'\beta_0)\right)\right|\right] a_1'\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)(Y_i-X_i'\beta_0)X_iX_i'\right]a_1\\ &\lesssim K_n\|\hat{\beta}-\beta_0\|_2^2 a_1'\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)(Y_i-X_i'\beta_0)X_iX_i'\right]a_1.\end{aligned}$$

Following the same steps as above, we obtain that $\|\mathbf{H}_{3,n}\|_{sp(\hat{s}_\beta)} \xrightarrow{p} 0$. Finally, we have

$$|a_1'\mathbf{H}_{4,n}a_1| \leq \left[\sup_{i=1,\ldots,n}\left|(1-D_i)\left(h''(X_i'\tilde{\beta})-h''(X_i'\beta_0)\right)X_i'(\mu_0-\tilde{\mu})\right|\right] a_1'\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)X_iX_i'\right]a_1$$
$$\lesssim K_n^2\|\hat{\beta}-\beta_0\|_2^2\|\hat{\mu}-\mu_0\|_2^2 a_1'\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)X_iX_i'\right]a_1,$$

and similar arguments give $\|\mathbf{H}_{4,n}\|_{sp(\hat{s}_\beta)} \xrightarrow{p} 0$. As a consequence, the first element of the right-hand side of (B.3) is bounded:

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(1-D_i)h''(X_i'\tilde{\beta})(Y_i-X_i'\tilde{\mu})X_iX_i'\right\|_{sp(\hat{s}_\beta)} \lesssim 1.$$

Now, decompose the second element of the right-hand side of (B.3) into two parts:

$$\frac{1}{n}\sum_{i=1}^{n}(1-D_i)h'(X_i'\tilde{\beta})X_iX_i' = \mathbf{H'}_{1,n}+\mathbf{H'}_{2,n},$$

where:

$$\mathbf{H'}_{1,n} = \frac{1}{n}\sum_{i=1}^{n}(1-D_i)h'(X_i'\beta_0)X_iX_i'$$
$$\mathbf{H'}_{2,n} = \frac{1}{n}\sum_{i=1}^{n}(1-D_i)\left[h'(X_i'\tilde{\beta})-h'(X_i'\beta_0)\right]X_iX_i'.$$

Using the inequality $uv \leq u^2+v^2/4$, we have that:

$$a_1'\mathbf{H'}_{1,n}a_2 = \frac{1}{n}\sum_{i=1}^{n}(1-D_i)\sqrt{|h'(X_i'\beta_0)|}(a_1'X_i)\sqrt{|h'(X_i'\beta_0)|}(X_i'a_2)$$
$$\leq a_1'\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_1)|h'(X_i'\beta_0)|X_iX_i'\right]a_1+\frac{1}{4}a_2'\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_1)|h'(X_i'\beta_0)|X_iX_i'\right]a_2.$$

Assumption 2.5, Lemma 1 and Supplemental Appendix C in Belloni and Chernozhukov (2013) imply that

$$\sup_{\substack{\|a\|_2>0\\ \|a_1\|_0\leq\hat{s}_\beta\\ \|a_2\|_0\leq\hat{s}_\mu}} a_1'\mathbf{H'}_{1,n}a_2 \lesssim 1.$$

Next, turn to $\mathbf{H}'_{2,n}$. Using a similar reasoning as above, we obtain:

$$a_1'\mathbf{H}'_{2,n}a_2 \leq \left[\sup_{i=1,\dots,n}\left|(1-D_i)\left(h'(X_i'\tilde{\beta})-h'(X_i'\beta_0)\right)\right|\right] a_1'\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)X_iX_i'\right]a_2$$
$$\lesssim K_n\|\hat{\beta}-\beta_0\|_2^2\left(a_1'\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)X_iX_i'\right]a_1+\frac{1}{4}a_2'\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)X_iX_i'\right]a_2\right).$$

Again, we find that

$$\sup_{\substack{\|a\|_2>0\\ \|a_1\|_0\leq\hat{s}_\beta\\ \|a_2\|_0\leq\hat{s}_\mu}} a_1'\mathbf{H}'_{2,n}a_2 \xrightarrow{p} 0.$$

All in all, coming back to equation (B.3), we can conclude that indeed:

$$\sup_{\substack{\|a\|_2>0\\ \|a_1\|_0\leq\hat{s}_\beta\\ \|a_2\|_0\leq\hat{s}_\mu}} \frac{a'\mathbb{E}_n\partial_\eta\partial_{\eta'}g(Z_i,\theta_0,\tilde{\eta})a}{\|a\|_2^2}\lesssim 1,$$

which proves that $I_3 \xrightarrow{p} 0$.

Conclusion: As a consequence of the previous steps:

$$\frac{\pi_0}{\sqrt{\mathbb{E}\left(g(Z,\theta_0,\eta_0)^2\right)}}\sqrt{n}(\hat{\theta}-\theta_0)\xrightarrow{d}\mathcal{N}(0,1),$$
$$\hat{\sigma}^{-1}\sqrt{n}(\hat{\theta}-\theta_0)\xrightarrow{d}\mathcal{N}(0,1),$$

with $\hat{\sigma}^2 := \mathbb{E}_n\left(g(Z_i,\hat{\theta},\hat{\eta})^2\right)/\mathbb{E}_n(D_i)^2$, a consistent estimator of the variance under Assumptions 2.4 and Theorem 2.3. □

Theorem 2.3 (Nuisance Parameter Estimation) *Under Assumptions 2.3-2.6, the nuisance parameters estimators defined in equations* (2.7) *and* (2.9) *satisfy, with probability tending to one,*

$$\|\hat{\beta}\|_0 \lesssim s_\beta, \tag{B.4}$$
$$\|\hat{\beta}-\beta_0\|_1 \lesssim s_\beta\sqrt{\log(p)/n}, \tag{B.5}$$
$$\|\hat{\beta}-\beta_0\|_2 \lesssim \sqrt{s_\beta\log(p)/n}, \tag{B.6}$$
$$\|\hat{\mu}\|_0 \lesssim s_\mu, \tag{B.7}$$
$$\|\hat{\mu}-\mu_0\|_1 \lesssim s_\mu\sqrt{\log(p)/n}, \tag{B.8}$$
$$\|\hat{\mu}-\mu_0\|_2 \lesssim \sqrt{s_\mu\log(p)/n}. \tag{B.9}$$

Proof of Theorem 2.3: The proof is divided into two parts, one for each of the nuisance parameters.

Part A: Verification of (B.4)-(B.6) *for* $\widehat{\beta}$.

Recall that $\hat{\beta}$ is defined as:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (1 - D_i) H(X_i'\beta) - D_i X_i'\beta + \lambda^d \sum_{j=1}^{p} \psi_{d,j} |\beta_j|, \tag{B.10}$$

with ideal penalty loadings satisfying Assumption 2.6. Let Ψ_d be the diagonal matrix with diagonal terms $\psi_{d,1}, \ldots, \psi_{d,p}$. We let $S_0 := \{j : \beta_{0j} \neq 0\}$, and recall that $\text{Card}(S_0) = \|\beta_0\|_0 \leq s_\beta$.

Step A.1: Concentration Inequality.

We first bound the sup-norm of the gradient of the objective function. Recall that for $p+1 \leq j \leq 2p$, $U_{i,j} = [(1 - D_i)h(X_i'\beta_0) - D_i] X_{i,j}$. Then let $\mathcal{S}_j := \sum_{i=1}^n U_{i,j}/\psi_{d,j}$ and the event

$$\mathcal{B}_{\lambda^d} := \left\{ \frac{1}{n} \max_{p+1 \leq j \leq 2p} \left| \sum_{i=1}^{n} \frac{U_{i,j}}{\psi_{d,j}} \right| \leq \frac{\lambda^d}{c} \right\}.$$

$(X_i, D_i)_{i=1}^n$ is a sequence of i.i.d. random vectors. By construction, $\mathbb{E}(U_{i,j}) = 0$. Moreover, by Assumptions 2.3 and 2.4, $\mathbb{E}(|U_{i,j}|^3) \leq +\infty$. Then, by Assumptions 2.3 and 2.6 and Lemma 5 in Belloni et al. (2012),

$$\begin{aligned}
\mathbb{P}\left(\mathcal{B}_{\lambda^d}^C\right) &= \mathbb{P}\left(\frac{c}{\sqrt{n}} \max_{p+1 \leq j \leq 2p} |\mathcal{S}_j| > c\Phi^{-1}(1 - \gamma/2p)/\sqrt{n} \right) \\
&= \mathbb{P}\left(\max_{p+1 \leq j \leq 2p} |\mathcal{S}_j| > \Phi^{-1}(1 - \gamma/2p) \right) \\
&\to 0.
\end{aligned}$$

Step A.2: Weighted Restricted Eigenvalue on Empirical Gram matrix.

This step links the restricted eigenvalue assumption on the theoretical Gram matrix Σ to one on its empirical counterpart $\hat{\Sigma}$. First, notice that Assumption 2.5 implies a weaker one because $S \subset S_{01}$:

$$\bar{\kappa}_\alpha^2(\Sigma) := \min_{\substack{S \subset \{1,\ldots,p\} \\ |S|_0 \leq s}} \min_{\delta \in \mathcal{C}[S,\alpha]} \frac{\delta'\Sigma\delta}{\|\delta_S\|_2^2} > 0.$$

We then show that Assumption 2.5 for Σ implies a restricted eigenvalue condition for $\hat{\Sigma}$. Set $\alpha = c_0$ and $\varepsilon := (1 - 7h\sqrt{(p + 2\log(2/\mu))/n})$ for some $\mu \to 0$, $\log(\mu) = o(n)$. By Theorem 3.1 in Oliveira (2016), we have, for any $\delta \in \mathbb{R}^p$, that $\delta'\hat{\Sigma}\delta \geq (1 - \varepsilon)\delta'\Sigma\delta$ with

probability tending to one. Then, by Assumption 2.5, with probability tending to one,

$$\min_{\substack{J\subset\{1,\ldots,p\}\\|J|_0\leq s}}\min_{\delta\in\mathcal{C}[J,\alpha]}\frac{\delta'\hat{\Sigma}\delta}{\|\delta_{J_{01}}\|_2^2}\geq(1-\varepsilon)\min_{\substack{J\subset\{1,\ldots,p\}\\|J|_0\leq s}}\min_{\delta\in\mathcal{C}[J,\alpha]}\frac{\delta'\Sigma\delta}{\|\delta_{J_{01}}\|_2^2}\geq(1-\varepsilon)\bar{\kappa}_\alpha^2(\Sigma).$$

Hence, $\kappa_\alpha(\hat{\Sigma})>0$. Secondly, notice that $\|\Psi_d\delta_{J_{01}}\|_2\leq\bar{\psi}\|\delta_{J_{01}}\|_2$. Consequently:

$$\frac{\sqrt{\delta'\hat{\Sigma}\delta}}{\|\Psi_d\delta_{J_{01}}\|_2}\geq\frac{1}{\bar{\psi}}\frac{\sqrt{\delta'\hat{\Sigma}\delta}}{\|\delta_{J_{01}}\|_2}.$$

Moreover, $\Psi_d\delta\in\mathcal{C}[J,\alpha]$ implies by Assumption 2.3 that $\delta\in\mathcal{C}[J,c_\psi\alpha]$. Then with probability tending to one,

$$\bar{\kappa}_\alpha(\hat{\Sigma})\geq\frac{\sqrt{1-\varepsilon}}{\bar{\psi}}\bar{\kappa}_{c_\psi\alpha}(\Sigma)>0.$$

Second, by the Cauchy-Schwarz inequality $\|\delta\|_1\leq\sqrt{\|\delta\|_0}\|\delta\|_2$. Hence, we have, with probability tending to one,

$$\tilde{\kappa}_\alpha(\hat{\Sigma}):=\min_{\substack{\Psi_d\delta\in\mathcal{C}[J,\alpha]\\|J|_0\leq s}}\sqrt{s}\frac{\sqrt{\delta'\hat{\Sigma}\delta}}{\|\Psi_d\delta_J\|_1}>0.$$

Step A.3: Basic Inequality.

We prove that with probability tending to one, $\hat{\beta}$ satisfies:

$$\underline{\mathrm{M}}(\hat{\beta}-\beta_0)'\hat{\Sigma}(\hat{\beta}-\beta_0)\leq 2\lambda^d\left(\|\Psi_d\beta_0\|_1-\|\Psi_d\hat{\beta}\|_1\right)+\frac{2\lambda^d}{c}\|\Psi_d(\hat{\beta}-\beta_0)\|_1,\qquad\text{(BASIC INEQUALITY)}$$

where $\underline{\mathrm{M}}$ is a lower bound on $\inf_{i=1,\ldots,n}h'(X_i'\hat{\beta})$ that does not depend on n.

By optimality of $\hat{\beta}$:

$$\frac{1}{n}\sum_{i=1}^n\gamma_{\hat{\beta}}^D(X_i,D_i)-\gamma_{\beta_0}^D(X_i,D_i)\leq\lambda^d\left(\|\Psi_d\beta_0\|_1-\|\Psi_d\hat{\beta}\|_1\right),$$

where $\gamma_\beta^D(X,D):=(1-D)H(X'\beta)-DX'\beta$. Subtract the inner product of the gradient $\bigtriangledown_\beta\gamma_{\beta_0}^D(X_i,D_i)$ and $\hat{\beta}-\beta_0$ on each side:

$$\frac{1}{n}\sum_{i=1}^n\gamma_{\hat{\beta}}^D(X_i,D_i)-\gamma_{\beta_0}^D(X_i,D_i)-\left((1-D_i)h(X_i'\beta_0)-D_i\right)(\hat{\beta}-\beta_0)'X_i\leq$$

$$\lambda^d\left(\|\Psi_d\beta_0\|_1-\|\Psi_d\hat{\beta}\|_1\right)-\frac{1}{n}\sum_{i=1}^n\left((1-D_i)h(X_i'\beta_0)-D_i\right)(\hat{\beta}-\beta_0)'X_i.$$

Now focus on the left-hand side of the equation. By Lemma 2.2, there exists $0 \leq t \leq 1$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\gamma_{\hat{\beta}}^{D}(X_i, D_i) - \gamma_{\beta_0}^{D}(X_i, D_i) - ((1-D_i)h(X_i'\beta_0) - D_i)\,(\hat{\beta}-\beta_0)'X_i =$$
$$\frac{1}{2}(\hat{\beta}-\beta_0)'\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)X_iX_i'h'(X_i'\tilde{\beta})\right](\hat{\beta}-\beta_0),$$

with $\tilde{\beta} := t\hat{\beta} + (1-t)\beta_0$. Plug this into the equation at the beginning of this paragraph and use $|\sum_i a_i b_i| \leq \max_i |b_i| \sum_i |a_i|$ on the right-hand side. On the event $\mathcal{B}_{\lambda^d}$ that occurs with probability $1 - o(1)$:

$$\frac{1}{2}(\hat{\beta}-\beta_0)'\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)X_iX_i'h'(X_i'\tilde{\beta})\right](\hat{\beta}-\beta_0) \leq \tag{B.11}$$
$$\lambda^d\left(\|\Psi_d\beta_0\|_1 - \|\Psi_d\hat{\beta}\|_1\right) - \frac{1}{n}\sum_{i=1}^{n}\left((1-D_i)h(X_i'\beta_0) - D_i\right)(\hat{\beta}-\beta_0)'X_i \leq$$
$$\lambda^d\left(\|\Psi_d\beta_0\|_1 - \|\Psi_d\hat{\beta}\|_1\right) + \|\Psi_d(\hat{\beta}-\beta_0)\|_1 \max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{X_{i,j}}{\psi_{d,j}}\left((1-D_i)h(X_i'\beta_0) - D_i\right)\right| \leq$$
$$\lambda^d\left(\|\Psi_d\beta_0\|_1 - \|\Psi_d\hat{\beta}\|_1\right) + \frac{\lambda^d}{c}\|\Psi_d(\hat{\beta}-\beta_0)\|_1.$$

The left-hand side is non-negative. Hence, provided that $c > 1$, we have, under $\mathcal{B}_{\lambda^d}$,

$$\begin{aligned}
0 &\leq \lambda^d\left(\|\Psi_d\beta_0\|_1 - \|\Psi_d\hat{\beta}\|_1\right) + \lambda^d/c\|\Psi_d(\hat{\beta}-\beta_0)\|_1 \\
\lambda^d\|\Psi_d\hat{\beta}\|_1 &\leq \lambda^d\|\Psi_d\beta_0\|_1 + \lambda^d/c\left(\|\Psi_d\hat{\beta}\|_1 + \|\Psi_d\beta_0\|_1\right) \\
\|\Psi_d\hat{\beta}\|_1 &\leq c_0\|\Psi_d\beta_0\|_1 \\
\|\hat{\beta}\|_1 &\leq c_0\frac{\bar{\psi}}{\underline{\psi}}\|\beta_0\|_1 \leq c_0\frac{\bar{\psi}}{\underline{\psi}}C_\beta,
\end{aligned}$$

where the last inequality follows from by Assumptions 2.4, with $c_0 = (c+1)/(c-1)$ and $\|\beta_0\|_1 \leq C_\beta$. As a consequence, $\underline{\mathrm{M}} := \inf_{i=1,\ldots,n} h'(X_i'\hat{\beta})$ is either equal to the lower bound of h' or to $h'(-K_n c_0 \frac{\bar{\psi}}{\underline{\psi}} C_\beta)$. The proof follows with:

$$\underline{\mathrm{M}}(\hat{\beta}-\beta_0)'\hat{\Sigma}(\hat{\beta}-\beta_0) \leq (\hat{\beta}-\beta_0)'\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)X_iX_i'h'(X_i'\tilde{\beta})\right](\hat{\beta}-\beta_0),$$

which gives a lower bound for inequality B.11 and gives us the desired result.

Step A.4: Control of ℓ_1-error for $\hat{\beta}$.

We prove that with probability tending to one,

$$\left\|\hat{\beta}-\beta_0\right\|_1 \leq \frac{2c_0\lambda^d s_\beta}{\underline{\psi}\underline{\mathrm{M}}\tilde{\kappa}_{c_0}(\hat{\Sigma})^2}.$$

The first step of the proof seeks to bound $\|\Psi_d\beta_0\|_1-\|\Psi_d\hat{\beta}\|_1$. By the triangular inequality:

$$\|\Psi_d\beta_{0,S_0}\|_1-\|\Psi_d\hat{\beta}_{S_0}\|_1 \leq \|\Psi_d(\beta_{0,S_0}-\hat{\beta}_{S_0})\|_1.$$

Focusing on the other part and using $|a-b|\leq|a|+|b|$:

$$\begin{aligned}\|\Psi_d\beta_{0,S_0^C}\|_1-\|\Psi_d\hat{\beta}_{S_0^C}\|_1 &= 2\|\Psi_d\beta_{0,S_0^C}\|_1-\|\Psi_d\beta_{0,S_0^C}\|_1-\|\Psi_d\hat{\beta}_{S_0^C}\|_1\\ &\leq 2\|\Psi_d\beta_{0,S_0^C}\|_1-\|\Psi_d(\beta_{0,S_0^C}-\hat{\beta}_{S_0^C})\|_1\\ &\leq -\|\Psi_d(\beta_{0,S_0^C}-\hat{\beta}_{S_0^C})\|_1.\end{aligned}$$

The last inequality comes from $\|\beta_{0,S_0^C}\|_1=0$ and assumptions 2.3. Consequently:

$$\begin{aligned}&\lambda^d\|\Psi_d\beta_0\|_1-\lambda^d\|\Psi_d\hat{\beta}\|_1+\frac{\lambda^d}{c}\|\Psi_d(\hat{\beta}-\beta_0)\|_1 \leq\\ &\qquad\lambda^d\left(1+\frac{1}{c}\right)\|\Psi_d(\hat{\beta}_{S_0}-\beta_{0,S_0})\|_1-\lambda^d\left(1-\frac{1}{c}\right)\|\Psi_d(\hat{\beta}_{S_0^C}-\beta_{0,S_0^C})\|_1.\end{aligned}$$

On $\mathcal{B}_{\lambda^d}$, by the basic inequality:

$$\underline{\mathrm{M}}(\hat{\beta}-\beta_0)'\hat{\Sigma}(\hat{\beta}-\beta_0) \leq 2\lambda^d\left(\|\Psi_d\beta_0\|_1-\|\Psi_d\hat{\beta}\|_1+\frac{1}{c}\|\Psi_d(\hat{\beta}-\beta_0)\|_1\right),$$

so :

$$\begin{aligned}&(\hat{\beta}-\beta_0)'\hat{\Sigma}(\hat{\beta}-\beta_0) \leq\\ &\frac{2}{\underline{\mathrm{M}}}\lambda^d\left[\left(1+\frac{1}{c}\right)\|\Psi_d(\hat{\beta}_{S_0}-\beta_{0,S_0})\|_1-\left(1-\frac{1}{c}\right)\|\Psi_d(\hat{\beta}_{S_0^C}-\beta_{0,S_0^C})\|_1\right].\end{aligned} \tag{B.12}$$

Using equation B.12 and $c>1$, it follows that:

$$(\hat{\beta}-\beta_0)'\hat{\Sigma}(\hat{\beta}-\beta_0)+\frac{2\lambda^d}{\underline{\mathrm{M}}}\left(1-\frac{1}{c}\right)\left\|\Psi_d(\hat{\beta}-\beta_0)\right\|_1 \leq \frac{4\lambda^d}{\underline{\mathrm{M}}}\left\|\Psi_d(\hat{\beta}_{S_0}-\beta_{0,S_0})\right\|_1.$$

From inequality B.12, because $(\hat{\beta}-\beta_0)'\hat{\Sigma}(\hat{\beta}-\beta_0)\geq 0$, notice that we have a cone condition $\Psi_d(\hat{\beta}-\beta_0)\in\mathcal{C}[S_0,c_0]$. So, we can use Step A.2 :

$$(\hat{\beta}-\beta_0)'\hat{\Sigma}(\hat{\beta}-\beta_0)+\frac{2\lambda^d}{\underline{\mathrm{M}}}\left(1-\frac{1}{c}\right)\left\|\Psi_d(\hat{\beta}-\beta_0)\right\|_1 \leq \frac{4\lambda^d\sqrt{s_\beta}}{\underline{\mathrm{M}}}\frac{\sqrt{(\hat{\beta}-\beta_0)'\hat{\Sigma}(\hat{\beta}-\beta_0)}}{\tilde{\kappa}_{c_0}(\hat{\Sigma})}.$$

Using the inequality $4uv \leq u^2 + 4v^2$:

$$(\hat{\beta} - \beta_0)'\hat{\Sigma}(\hat{\beta} - \beta_0) + \frac{2\lambda^d}{\underline{\mathrm{M}}}\left(1 - \frac{1}{c}\right)\left\|\Psi_d(\hat{\beta} - \beta_0)\right\|_1 \leq (\hat{\beta} - \beta_0)'\hat{\Sigma}(\hat{\beta} - \beta_0) + \frac{4\lambda^{d2} s_\beta}{\underline{\mathrm{M}}^2 \tilde{\kappa}_{c_0}(\hat{\Sigma})^2}.$$

Consequently:

$$\left\|\Psi_d(\hat{\beta} - \beta_0)\right\|_1 \leq \left(\frac{c}{c-1}\right)\frac{2\lambda^d s_\beta}{\underline{\mathrm{M}}\tilde{\kappa}_{c_0}(\hat{\Sigma})^2} \leq \frac{2c_0\lambda^d s_\beta}{\underline{\mathrm{M}}\tilde{\kappa}_{c_0}(\hat{\Sigma})^2},$$

where the last line results from $c/(c-1) < c_0$.

Step A.5: Control of ℓ_2-error for $\hat{\beta}$.
We prove that with probability tending to one:

$$\|\hat{\beta} - \beta_0\|_2 \leq \left(1 + c_0\sqrt{\frac{s_\beta}{m}}\right)\frac{4\lambda^d\sqrt{s_\beta}}{\underline{\psi}\underline{\mathrm{M}}\kappa_{c_0}^2(\hat{\Sigma})}.$$

Because $m \geq s_\beta$, we obtain

$$\|\hat{\beta} - \beta_0\|_2 \lesssim \sqrt{\frac{s_\beta \log(p)}{n}}.$$

Using the cone condition $\Psi_d(\hat{\beta} - \beta_0) \in \mathcal{C}[S_0, c_0]$:

$$\begin{aligned}\left\|\Psi_d(\hat{\beta} - \beta_0)\right\|_1 &= \left\|\Psi_d(\hat{\beta}_{S_0} - \beta_{0,S_0})\right\|_1 + \left\|\Psi_d(\hat{\beta}_{S_0^C} - \beta_{0,S_0^C})\right\|_1 \\ &\leq (1 + c_0)\left\|\Psi_d(\hat{\beta}_{S_0} - \beta_{0,S_0})\right\|_1 \\ &\leq (1 + c_0)\sqrt{s_\beta}\left\|\Psi_d(\hat{\beta}_{S_0} - \beta_{0,S_0})\right\|_2.\end{aligned}$$

Denote $S_{01} = S_0 \cup S_1$, where S_1 is the set of indices corresponding to the m largest elements of $\Psi_d(\hat{\beta} - \beta_0)$ whose index is not in S_0. Notice that the k-th largest in absolute value element of $\Psi_d(\hat{\beta}_{S_0^C} - \beta_{0,S_0^C})$ satisfies: $k|\Psi_d(\hat{\beta}_{S_0^C} - \beta_{0,S_0^C})|_{(k)} \leq \left\|\Psi_d(\hat{\beta}_{S_0^C} - \beta_{0,S_0^C})\right\|_1$.

$$\left\|\Psi_d(\hat{\beta}_{S_{01}^C} - \beta_{0,S_{01}^C})\right\|_2^2 \leq \left\|\Psi_d(\hat{\beta}_{S_0^C} - \beta_{0,S_0^C})\right\|_1^2 \sum_{k \geq m+1} k^{-2} \leq \frac{1}{m}\left\|\Psi_d(\hat{\beta}_{S_0^C} - \beta_{0,S_0^C})\right\|_1^2.$$

From the above inequality and the cone condition $\Psi_d(\hat{\beta} - \beta_0) \in \mathcal{C}[S_0, c_0]$:

$$\begin{aligned}\left\|\Psi_d(\hat{\beta}_{S_{01}^C} - \beta_{0,S_{01}^C})\right\|_2 &\leq \frac{1}{\sqrt{m}}\left\|\Psi_d(\hat{\beta}_{S_0^C} - \beta_{0,S_0^C})\right\|_1 \\ &\leq \frac{c_0}{\sqrt{m}}\left\|\Psi_d(\hat{\beta}_{S_0} - \beta_{0,S_0})\right\|_1\end{aligned}$$

$$\leq c_0\sqrt{\frac{s_\beta}{m}}\left\|\Psi_d(\hat{\beta}_{S_0}-\beta_{0,S_0})\right\|_2$$
$$\leq c_0\sqrt{\frac{s_\beta}{m}}\left\|\Psi_d(\hat{\beta}_{S_{01}}-\beta_{0,S_{01}})\right\|_2.$$

This last inequality gives:

$$\left\|\Psi_d(\hat{\beta}-\beta_0)\right\|_2 \leq \left(1+c_0\sqrt{\frac{s_\beta}{m}}\right)\left\|\Psi_d(\hat{\beta}_{S_{01}}-\beta_{0,S_{01}})\right\|_2.$$

We want to find an upper bound of the term in the right-hand side above. Using equation (B.12) and $c>1$, we have:

$$(\hat{\beta}-\beta_0)'\hat{\Sigma}(\hat{\beta}-\beta_0) \leq \frac{4\lambda^d}{\underline{\mathrm{M}}}\left\|\Psi_d(\hat{\beta}_{S_0}-\beta_{0,S_0})\right\|_1$$
$$\leq \frac{4\lambda^d}{\underline{\mathrm{M}}}\left\|\Psi_d(\hat{\beta}_{S_{01}}-\beta_{0,S_{01}})\right\|_1$$
$$\leq \frac{4\lambda^d\sqrt{s_\beta}}{\underline{\mathrm{M}}}\left\|\Psi_d(\hat{\beta}_{S_{01}}-\beta_{0,S_{01}})\right\|_2.$$

Using Step A.2 to bound $(\hat{\beta}-\beta_0)'\hat{\Sigma}(\hat{\beta}-\beta_0)$ from below and the inequality above, we obtain:

$$\left\|\Psi_d(\hat{\beta}_{S_{01}}-\beta_{0,S_{01}})\right\|_2 \leq \frac{4\lambda^d\sqrt{s_\beta}}{\underline{\mathrm{M}}\kappa^2_{c_0}(\hat{\Sigma})}.$$

Thus, we obtain:

$$\left\|\Psi_d(\hat{\beta}-\beta_0)\right\|_2 \leq \left(1+c_0\sqrt{\frac{s_\beta}{m}}\right)\frac{4\lambda^d\sqrt{s_\beta}}{\underline{\mathrm{M}}\kappa^2_{c_0}(\hat{\Sigma})}.$$

Step A.6: Empirical sparsity for $\hat{\beta}$.

We prove that with probability tending to one $\|\hat{\beta}\|_0 \lesssim s_\beta$.

Denote $\hat{S} := \left\{j : \hat{\beta}_j \neq 0\right\}$, the set of indices which indicates non-zero coefficients in $\hat{\beta}$. From KKT optimality conditions, we have that:

$$\left|\frac{1}{n}\sum_{i=1}^n[(1-D_i)h(X_i'\hat{\beta})-D_i]\frac{X_{i,j}}{\psi_{d,j}}\right| = \lambda^d, \text{for all } j\in\hat{S}.$$

So:

$$\lambda^d\sqrt{\|\hat{\beta}\|_0} \leq \left\|\frac{1}{n}\sum_{i=1}^n[(1-D_i)h(X_i'\hat{\beta})-D_i](\Psi^{d-1}X_i)_{\hat{S}}\right\|_2$$

$$\leq \left\| \frac{1}{n} \sum_{i=1}^{n} [(1-D_i)h(X_i'\beta_0) - D_i](\Psi^{d-1}X_i)_{\hat{S}} \right\|_2 + \left\| \frac{1}{n} \sum_{i=1}^{n} (1-D_i)[h(X_i'\hat{\beta}) - h(X_i'\beta_0)](\Psi^{d-1}X_i)_{\hat{S}} \right\|_2 . \quad \text{(B.13)}$$

Using the definition of the event $\mathcal{B}_{\lambda^d}$, with probability tending to one:

$$\left\| \frac{1}{n} \sum_{i=1}^{n} [(1-D_i)h(X_i'\beta_0) - D_i](\Psi^{d-1}X_i)_{\hat{S}} \right\|_2 \leq \sqrt{\|\hat{\beta}\|_0} \left\| \frac{1}{n} \sum_{i=1}^{n} [(1-D_i)h(X_i'\beta_0) - D_i]\Psi^{d-1}X_i \right\|_\infty \leq \frac{\lambda^d}{c}\sqrt{\|\hat{\beta}\|_0}. \quad \text{(B.14)}$$

Define $\hat{m}_\beta := |\hat{S} \backslash S_0|$. For notational simplicity, let us introduce $\mathbf{X_0} = ((1-D_i)X_{i,j})_{1\leq i\leq n, 1\leq j\leq p}$ and $\mathbf{F_0} = \Big((1-D_i)[h(X_i'\hat{\beta}) - h(X_i'\beta_0)]\Big)_{1\leq i\leq n}$

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_{i=1}^{n} (1-D_i)[h(X_i'\hat{\beta}) - h(X_i'\beta_0)](\Psi^{d-1}X_i)_{\hat{S}} \right\|_2 \\
\leq &\frac{1}{\underline{\psi}} \sup_{\substack{\|\delta_{S_0^C}\|_0 \leq \hat{m}_\beta \\ \|\delta\|_2 \leq 1}} \left| \delta' \left(\frac{1}{n} \sum_{i=1}^{n} (1-D_i)[h(X_i'\hat{\beta}) - h(X_i'\beta_0)]X_i \right) \right| \\
\leq &\frac{1}{\underline{\psi}} \sup_{\substack{\|\delta_{S_0^C}\|_0 \leq \hat{m}_\beta \\ \|\delta\|_2 \leq 1}} \left| \frac{1}{n} \delta' \mathbf{X_0}' \mathbf{F_0} \right| \\
\leq &\frac{1}{\underline{\psi}} \sup_{\substack{\|\delta_{S_0^C}\|_0 \leq \hat{m}_\beta \\ \|\delta\|_2 \leq 1}} \left\| \frac{1}{\sqrt{n}} \delta' \mathbf{X_0}' \right\|_2 \left\| \frac{1}{\sqrt{n}} \mathbf{F_0} \right\|_2 \\
\leq &\frac{1}{\underline{\psi}} \left\| \frac{1}{\sqrt{n}} \mathbf{F_0} \right\|_2 \left\| \hat{\Sigma} \right\|_{sp(\hat{m}_\beta)},
\end{aligned} \quad \text{(B.15)}$$

Assumptions 2.4-2.5 imply that $\left\| \hat{\Sigma} \right\|_{sp(\hat{m}_\beta)}$ is bounded if $\hat{m} + s_\beta \leq s_\beta \log(n)$ (see Lemma 1 and Supplemental Appendix C in Belloni and Chernozhukov, 2013). Next, we have

$$\left\| \frac{1}{\sqrt{n}} \mathbf{F_0} \right\|_2 \leq C_{h'} \left\| \hat{\beta} - \beta_0 \right\|_2 \left\| \hat{\Sigma} \right\|_{sp(\hat{m}_\beta)}.$$

Combining (B.13)-(B.15), we obtain:

$$\lambda^d \left(1 - \frac{1}{c}\right) \sqrt{\|\hat{\beta}\|_0} \leq \frac{C_{h'}}{\underline{\psi}} \left\| \hat{\beta} - \beta_0 \right\|_2 \left\| \hat{\Sigma} \right\|^2_{sp(\hat{m}_\beta)},$$

which completes the proof in light of the result displayed in the previous step.

Part B: Verification of (B.4)-(B.6) *for* $\widehat{\mu}$.

Recall that $\widehat{\mu}$ is defined as:

$$\hat{\mu} = \arg\min_{\mu} \frac{1}{n} \sum_{i=1}^{n} (1 - D_i) h'(X_i'\hat{\beta}) \left(Y_i - X_i'\mu\right)^2 + \lambda^y \sum_{j=1}^{p} \psi_{y,j} |\mu_j|,$$

This estimator is a weighted version of the usual Lasso. The steps needed to achieve (B.4)-(B.6) for $\widehat{\mu}$ closely follow the ones from the previous subsection or can be found in Belloni et al. (2012) for example. For the sake of clarity we will state the points where the proof differs from the one before.

Step B.1: Concentration Inequality.

Another concentration inequality is needed to bound the sup-norm of the gradient of the objective function. Define $V_{i,j} := (1 - D_i) h'(X_i'\beta_0) \left[Y_i - X_i'\mu_0\right] X_{i,j}$, $\mathcal{S}_j' := \sum_{i=1}^{n} V_{ij}/\psi_{y,j}$ and the following event

$$\mathcal{B}'_{\lambda^y} := \left\{ \frac{2}{n} \max_{1 \leq j \leq p} \left| \sum_{i=1}^{n} \frac{V_{ij}}{\psi_{y,j}} \right| \leq \frac{\lambda^y}{c} \right\}.$$

$(Y_i, X_i, D_i)_{i=1}^n$ is a sequence of i.i.d. random vectors. By construction, $\mathbb{E}(V_{ij}) = 0$. Moreover, by Assumptions 2.3 and 2.4, the variables V_{ij} have finite third-order moments, $\mathbb{E}(|V_{ij}|^3) \leq +\infty$. Then, by Assumptions 2.3 and 2.6, and Lemma 5 in Belloni et al. (2012),

$$\begin{aligned}
\mathbb{P}\left(\mathcal{B}_{\lambda}^{\prime C}\right) &= \mathbb{P}\left(\frac{c}{\sqrt{n}} \max_{1 \leq j \leq p} |\mathcal{S}_j'| > c\Phi^{-1}(1 - \gamma/2p)/\sqrt{n} \right) \\
&= \mathbb{P}\left(\max_{1 \leq j \leq p} |\mathcal{S}_j'| > \Phi^{-1}(1 - \gamma/2p) \right) \\
&\to 0.
\end{aligned}$$

Step B.2: Control of ℓ_1 *and* ℓ_2*-errors for* $\hat{\mu}$.

We prove that with probability tending to one:

$$\|\Psi_y(\hat{\mu} - \mu_0)\|_1 \leq \frac{c_0 \lambda^y s_\mu}{\left(\underline{\mathrm{M}} - C_{h'} K_n \|\hat{\beta} - \beta_0\|_1\right) \bar{\kappa}_\alpha^2(\hat{\Sigma}) - 2\sqrt{\alpha_n}}.$$

This step is complicated because the empirical loss function for $\hat{\mu}$ depends on $\hat{\beta}$ rather than β_0. Denote $\gamma^Y_{\beta,\mu}(Z_i) := (1 - D_i) h'(X_i'\beta) \left(Y_i - X_i'\mu\right)^2$. Use the decomposition:

$$\frac{1}{n} \sum_{i=1}^{n} \gamma^Y_{\beta_0,\hat{\mu}}(Z_i) - \gamma^Y_{\beta_0,\mu_0}(Z_i) = R_n + \frac{1}{n} \sum_{i=1}^{n} \gamma^Y_{\hat{\beta},\hat{\mu}}(Z_i) - \gamma^Y_{\hat{\beta},\mu_0}(Z_i),$$

with $R_n := (1/n)\sum_{i=1}^n \gamma^Y_{\hat\beta_0,\hat\mu}(Z_i) - \gamma^Y_{\hat\beta,\hat\mu}(Z_i) - \gamma^Y_{\beta_0,\mu_0}(Z_i) + \gamma^Y_{\hat\beta,\mu_0}(Z_i)$. Rewrite this remainder term as $R_n = R_n^1 + R_n^2$, with:

$$R_n^1 := \frac{1}{n}\sum_{i=1}^n (1-D_i)\left[h'(X_i'\beta_0) - h'(X_i'\hat\beta)\right]\left[X_i'(\mu_0-\hat\mu)\right]^2,$$
$$R_n^2 := \left[\frac{2}{n}\sum_{i=1}^n (1-D_i)\left[h'(X_i'\beta_0) - h'(X_i'\hat\beta)\right](Y_i - X_i'\mu_0)X_i\right]'(\mu_0-\hat\mu).$$

With probability tending to one, $[\min_i X_i'\widehat\beta, \max_i X_i'\widehat\beta] \subset K$ compact. Then, because h' is Lipschitz on this compact (with Lipschitz constant $C_{h'}$, say),

$$|h'(X_i'\widehat\beta) - h'(X_i'\beta_0)| \le C_{h'}|X_i'(\widehat\beta-\beta_0)| \le C_{h'}K_n\|\hat\beta-\beta_0\|_1,$$

so we have:

$$|R_n^1| \le C_{h'}K_n\|\hat\beta-\beta_0\|_1(\hat\mu-\mu_0)'\hat\Sigma(\hat\mu-\mu_0). \tag{B.16}$$

The second term satisfies (see Step A.6 for more details):

$$|R_n^2| < 2C_{h'}\|\hat\mu-\mu_0\|_2\|\hat\beta-\beta_0\|_2\left\|\frac{1}{n}\sum_{i=1}^n (1-D_i)(Y_i-X_i'\mu_0)^2X_iX_i'\right\|_{sp(\hat m_\beta)}. \tag{B.17}$$

The sparse norm of the matrix in the term of right-hand side is bounded using Assumptions 2.5 and 2.4 and (Lemma 1 and Supplemental Appendix C in Belloni and Chernozhukov, 2013). Because $\hat\mu$ is the minimizer of the empirical loss function we obtain:

$$\frac{1}{n}\sum_{i=1}^n \gamma^Y_{\hat\beta,\hat\mu}(Z_i) - \gamma^Y_{\hat\beta,\mu_0}(Z_i) \le \lambda^y\left(\|\Psi_y\mu_0\|_1 - \|\Psi_y\hat\mu\|_1\right). \tag{B.18}$$

On the other hand, we have:

$$\frac{1}{n}\sum_{i=1}^n \gamma^Y_{\beta_0,\hat\mu}(Z_i) - \gamma^Y_{\beta_0,\mu_0}(Z_i) + 2(1-D_i)h'(X_i'\beta_0)(Y_i-X_i'\mu_0)X_i'(\hat\mu-\mu_0) =$$
$$(\hat\mu-\mu_0)'\left(\frac{1}{n}\sum_{i=1}^n (1-D_i)h'(X_i'\beta_0)X_iX_i'\right)(\hat\mu-\mu_0).$$

Combine the previous equality with the inequalities B.16, B.17, B.18, and the concentration inequality for the gradient of the current loss function in Step A.4 to obtain:

$$(\hat\mu-\mu_0)'\left(\frac{1}{n}\sum_{i=1}^n (1-D_i)h'(X_i'\beta_0)X_iX_i'\right)(\hat\mu-\mu_0) \le \lambda^y\left(\|\Psi_y\mu_0\|_1 - \|\Psi_y\hat\mu\|_1\right)$$
$$+ C_{h'}K_n\|\hat\beta-\beta_0\|_1(\hat\mu-\mu_0)'\hat\Sigma(\hat\mu-\mu_0)$$

$$+ 2C_{h'}\|\hat{\mu}-\mu_0\|_2\|\hat{\beta}-\beta_0\|_2 \left\|\frac{1}{n}\sum_{i=1}^{n}(1-D_i)\left(Y_i - X_i'\mu_0\right)^2 X_iX_i'\right\|_{sp(\hat{m}_\beta)}$$
$$+ \frac{\lambda^y}{c}\|\hat{\mu}-\mu_0\|_1.$$

The two terms in the middle of the right-hand side are what separates us from the classical case. Denote:

$$L_n := \left\|\frac{1}{n}\sum_{i=1}^{n}(1-D_i)\left(Y_i - X_i'\mu_0\right)^2 X_iX_i'\right\|_{sp(\hat{m}_\beta)} \|\hat{\beta}-\beta_0\|_2$$

Consider a series $(\alpha_n)_n$ such that $\alpha_n \to 0$ and $\sqrt{n}\alpha_n/s_\beta \log(p) \to +\infty$. We need to consider two cases.

1. $\|\hat{\mu}-\mu_0\|_2^2 \leq L_n^2/\alpha_n$. This is the trivial case where we obtain $\|\hat{\mu}-\mu_0\|_2^2 \lesssim s_\beta log(p)/(n\alpha_n) = o_P(1/n)$, and the ℓ_1-rate follows by Cauchy-Schwarz.

2. $\|\hat{\mu}-\mu_0\|_2^2 > L_n^2/\alpha_n$. Consequently $2L_n\|\hat{\mu}-\mu_0\|_2 \leq 2\sqrt{\alpha_n}\|\hat{\mu}-\mu_0\|_2^2$. Reproducing the same reasoning as in Step A.4, the previous inequality becomes:

$$(\hat{\mu}-\mu_0)'\left[\left(\underline{\mathrm{M}} - C_{h'}K_n\|\hat{\beta}-\beta_0\|_1\right)\hat{\Sigma} - 2\sqrt{\alpha_n}I_p\right](\hat{\mu}-\mu_0)$$
$$\leq \lambda^y\left[\left(1+\frac{1}{c}\right)\|\Psi_y(\hat{\mu}_{S_0}-\mu_{0,S_0})\|_1 - \left(1-\frac{1}{c}\right)\|\Psi_y(\hat{\mu}_{S_0^C}-\mu_{0,S_0^C})\|_1 + 2c_1\bar{\psi}\sqrt{\frac{s_\mu^2}{n}}\right].$$

We obtain a cone condition, $\|\Psi_y(\hat{\mu}_{S_0^C}-\mu_{0,S_0^C})\|_1 \leq c_0\|\Psi_y(\hat{\mu}_{S_0}-\mu_{0,S_0})\|_1$, which allows to use the restricted eigenvalue condition above. For n sufficiently large so we have $2\sqrt{\alpha_n} < \left(\underline{\mathrm{M}} - C_{h'}K_n\|\hat{\beta}-\beta_0\|_1\right)\bar{\kappa}_\alpha^2(\hat{\Sigma})$, the following restricted eigenvalue condition holds:

$$\min_{\substack{\Psi_y\delta\in\mathcal{C}[J,\alpha]\\ |J|_0\leq s}} \frac{\delta'\left[\left(\underline{\mathrm{M}} - C_{h'}K_n\|\hat{\beta}-\beta_0\|_1\right)\hat{\Sigma} - 2\sqrt{\alpha_n}I_p\right]\delta}{\|\Psi_y\delta_J\|_2^2} >$$
$$\left(\underline{\mathrm{M}} - C_{h'}K_n\|\hat{\beta}-\beta_0\|_1\right)\bar{\kappa}_\alpha^2(\hat{\Sigma}) - 2\sqrt{\alpha_n} > 0,$$

and a similar condition with the ℓ_1-norm at the denominator also holds by Cauchy-Schwarz inequality. Following the same path as in Step A.4, we arrive at:

$$\|\Psi_y(\hat{\mu}-\mu_0)\|_1 \leq \frac{c_0\lambda^y s_\mu}{\left(\underline{\mathrm{M}} - C_{h'}K_n\|\hat{\beta}-\beta_0\|_1\right)\bar{\kappa}_\alpha^2(\hat{\Sigma}) - 2\sqrt{\alpha_n}}.$$

Once that step is completed, the proof for $\hat{\mu}$ follows the same steps as for $\hat{\beta}$.

Step B.3: Empirical sparsity for $\hat{\mu}$.

We prove that with probability tending to one:

$$\|\hat{\mu}\|_0 \lesssim s_\mu.$$

Using the same reasoning as in Step A.6, we have, this time defining $\hat{S} := \{j : \hat{\mu}_j \neq 0\}$:

$$\lambda^y \sqrt{\|\hat{\mu}\|_0} \leq \left\| \frac{2}{n} \sum_{i=1}^{n} (1 - D_i) h'(X_i'\hat{\beta})(Y_i - X_i'\hat{\mu})(\Psi_y^{-1} X_i)_{\hat{S}} \right\|_2.$$

Decompose the individual contribution to the gradient of the loss function in four parts:

$$\begin{aligned}
(1 - D_i) h'(X_i'\hat{\beta})(Y_i - X_i'\hat{\mu})\Psi_y^{-1} X_i = {} & (1 - D_i) h'(X_i'\beta_0)(Y_i - X_i'\mu_0)\Psi_y^{-1} X_i \\
& + (1 - D_i) h'(X_i'\beta_0) X_i'(\mu_0 - \hat{\mu})\Psi_y^{-1} X_i \\
& + (1 - D_i)(h'(X_i'\hat{\beta}) - h'(X_i'\beta_0))(Y_i - X_i'\mu_0)\Psi_y^{-1} X_i \\
& + (1 - D_i)(h'(X_i'\hat{\beta}) - h'(X_i'\beta_0)) X_i'(\mu_0 - \hat{\mu})\Psi_y^{-1} X_i.
\end{aligned}$$

On $\mathcal{B}'_{\lambda^y}$, the ℓ_2-norm of the first part obeys:

$$\left\| \frac{2}{n} \sum_{i=1}^{n} (1 - D_i) h'(X_i'\beta_0)(Y_i - X_i'\mu_0)(\Psi_y^{-1} X_i)_{\hat{S}} \right\|_2 \leq \frac{\lambda^y}{c} \sqrt{\|\hat{\mu}\|_0}.$$

Define:

$$\begin{aligned}
\mathbf{G_1} &:= ((1 - D_i) h'(X_i'\beta_0) X_i'(\mu_0 - \hat{\mu}))_{1 \leq i \leq n} \\
\mathbf{G_2} &:= \Big((1 - D_i)(h'(X_i'\hat{\beta}) - h'(X_i'\beta_0))(Y_i - X_i'\mu_0) X_i \Big)_{1 \leq i \leq n} \\
\mathbf{G_3} &:= \Big((1 - D_i)(h'(X_i'\hat{\beta}) - h'(X_i'\beta_0)) X_i'(\mu_0 - \hat{\mu}) \Big)_{1 \leq i \leq n},
\end{aligned}$$

and similarly, $T_j := \frac{2}{n} \sum_{i=1}^{n} \mathbf{G}_{j,i} (\Psi_y^{-1} X_i)_{\hat{S}}$, for $j = 1, 2, 3$. Any of the T_j satisfies:

$$\begin{aligned}
\|T_j\|_2 &\leq \frac{1}{\underline{\psi}} \sup_{\substack{\left\|\delta_{S_0^C}\right\|_0 \leq \hat{m}_\mu \\ \|\delta\|_2 \leq 1}} \left| \frac{2}{n} \delta' \mathbf{X_0}' \mathbf{G}_j \right| \\
&\leq \frac{2}{\underline{\psi}} \left\| \frac{\mathbf{G}_j}{\sqrt{n}} \right\|_2 \left\| \hat{\Sigma} \right\|_{sp(\hat{m}_\mu)}.
\end{aligned}$$

We now need to bound $\|\mathbf{G}_j/\sqrt{n}\|_2$ for $j = 1, 2, 3$. Firstly, using Holder inequality:

$$\left\| \frac{1}{\sqrt{n}} \mathbf{G_1} \right\|_2 \leq \|\hat{\mu} - \mu_0\|_2 \left\| \frac{1}{n} \sum_{i=1}^{n} (1 - D_i) h'(X_i'\beta_0)^2 X_i X_i' \right\|_{sp(\hat{m}_\mu)},$$

where the sparse norm of the matrix in the term of left-hand side is bounded using Assumptions 2.5 and 2.4 and (Lemma 1 and Supplemental Appendix C in Belloni and Chernozhukov, 2013). Secondly:

$$\left\|\frac{1}{\sqrt{n}}\mathbf{G_2}\right\|_2 \leq C_{h'}\left\|\hat{\beta}-\beta_0\right\|_2\left\|\frac{1}{n}\sum_{i=1}^{n}(1-D_i)(Y_i-X_i'\mu_0)^2X_iX_i'\right\|_{sp(\hat{m}_\beta)},$$

and the sparse norm of the matrix in the term of left-hand side is bounded using similar arguments as the line above. Thirdly, using Cauchy-Schwarz inequality on the third line:

$$\begin{aligned}
\left\|\frac{1}{\sqrt{n}}\mathbf{G_3}\right\|_2 &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(1-D_i)\left[h'(X_i'\hat{\beta})-h'(X_i'\beta_0)\right]^2(\hat{\mu}-\mu_0)'X_iX_i'(\hat{\mu}-\mu_0)} \\
&\leq C_{h'}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(1-D_i)\left(\hat{\beta}-\beta_0\right)'X_iX_i'\left(\hat{\beta}-\beta_0\right)(\hat{\mu}-\mu_0)'X_iX_i'(\hat{\mu}-\mu_0)} \\
&\leq C_{h'}\left(\sqrt{\left(\hat{\beta}-\beta_0\right)'\hat{\Sigma}\left(\hat{\beta}-\beta_0\right)}\sqrt{(\hat{\mu}-\mu_0)'\hat{\Sigma}(\hat{\mu}-\mu_0)}\right)^{1/2} \\
&\leq C_{h'}\left(\left\|\hat{\beta}-\beta_0\right\|_2\|\hat{\mu}-\mu_0\|_2\left\|\hat{\Sigma}\right\|_{sp(\hat{m}_\mu)}\left\|\hat{\Sigma}\right\|_{sp(\hat{m}_\beta)}\right)^{1/2}.
\end{aligned}$$

Notice that:

$$\begin{aligned}
\lambda^y\left(1-\frac{1}{c}\right)\sqrt{\|\hat{\mu}\|_0} &\leq \frac{2}{\underline{\psi}}\left\|\hat{\Sigma}\right\|_{sp(\hat{m}_\mu)}\sum_{j=1}^{3}\left\|\frac{\mathbf{G}_j}{\sqrt{n}}\right\|_2 \\
&\leq \frac{2}{\underline{\psi}}\left\|\hat{\Sigma}\right\|_{sp(\hat{m}_\mu)}\left(\|\hat{\mu}-\mu_0\|_2\left\|\frac{1}{n}\sum_{i=1}^{n}(1-D_i)h'(X_i'\beta_0)^2X_iX_i'\right\|_{sp(\hat{m}_\mu)}\right. \\
&+ C_{h'}\left\|\hat{\beta}-\beta_0\right\|_2\left\|\frac{1}{n}\sum_{i=1}^{n}(1-D_i)(Y_i-X_i'\mu_0)^2X_iX_i'\right\|_{sp(\hat{m}_\beta)} \\
&\left.+ C_{h'}\sqrt{\left\|\hat{\beta}-\beta_0\right\|_2\|\hat{\mu}-\mu_0\|_2\left\|\hat{\Sigma}\right\|_{sp(\hat{m}_\mu)}\left\|\hat{\Sigma}\right\|_{sp(\hat{m}_\beta)}}\right),
\end{aligned}$$

which proves the result in light of Steps A.4, A.5, B.2 and Assumption 2.3. □

Lemma 2.2 (A Taylor Expansion Lemma) *Assume that f is a C^2 function from $\mathbb{R}^p$ to $\mathbb{R}$. Then, for any $(x, x_0) \in (\mathbb{R}^p)^2$, there exists $t \in (0,1)$ such that*

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{1}{2}(x-x_0)'f''(x')(x-x_0),$$

with $x' = tx_0+(1-t)x$ and where $f'(x_0)$ and $f''(x')$ denote the gradient vector and hessian matrix of f taken at x_0 and x', respectively.

Proof of Lemma 2.2: Let h be a real function defined on $[0,1]$ by

$$h(t) = f(x) - f(tx_0 + (1-t)x) - f'(tx_0 + (1-t)x)(x - x_0)t - At^2$$

with A defined such that $h(1) = 0$. Note that we also have $h(0) = 0$. Then, by Rolle's theorem, since h is C^1, there exists $t \in (0,1)$ such that $h'(t) = 0$. This yields, with $x' = tx_0 + (1-t)x$,

$$-f'(x')(x_0 - x) - f'(x')(x - x_0) + (x - x_0)'f''(x')(x - x_0)t - 2At = 0.$$

Thus, because $t > 0$,

$$A = \frac{1}{2}(x - x_0)'f''(x')(x - x_0).$$

The result follows using $h(1) = 0$. □

Chapter 3

A Penalized Synthetic Control Estimator for Disaggregated Data

Joint work with Alberto Abadie.

Summary

Synthetic control methods are commonly applied in empirical research to estimate the effects of treatments or interventions of interest on aggregate outcomes. A synthetic control estimator compares the outcome of a treated unit to the outcome of a weighted average of untreated units that best resembles the characteristics of the treated unit before the intervention. When disaggregated data are available, constructing separate synthetic controls for each treated unit may help avoid interpolation biases. However, the problem of finding a synthetic control that best reproduces the characteristics of a treated unit may not have a unique solution. Multiplicity of solutions is a particularly daunting challenge in settings with disaggregated data, that is, when the sample includes many treated and untreated units. To address this challenge, we propose a synthetic control estimator that penalizes the pairwise discrepancies between the characteristics of the treated units and the characteristics of the units that contribute to their synthetic controls. The penalization parameter trades off pairwise matching discrepancies with respect to the characteristics of each unit in the synthetic control against matching discrepancies with respect to the characteristics of the synthetic control unit as a whole. We study the properties of this estimator and propose data driven choices of the penalization parameter.

1. Introduction

Synthetic control methods (Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015; Doudchenko and Imbens, 2016) are often applied to estimate the treatment effects of aggregate interventions (see, e.g., Kleven et al., 2013; Bohn et al., 2014a; Hackmann et al., 2015; Cunningham and Shah, 2018). Suppose we observe data for a unit that is affected by the treatment or intervention of interest, as well as data on a donor pool, that is, a set of untreated units that are available to approximate the outcome that would have been observed for the treated unit in the absence of the intervention. The idea behind synthetic controls is to match each unit exposed to the intervention or treatment of interest to a weighted average of the units in the donor pool that most closely resembles the characteristics of the treated unit before the intervention. Once a suitable synthetic control is selected, differences in outcomes between the treated unit and the synthetic control are taken as estimates of the effect of the treatment on the unit exposed to the intervention of interest.

The synthetic control method is akin to nearest neighbor matching estimators (Dehejia and Wahba, 2002; Abadie and Imbens, 2006; Imbens and Rubin, 2015) but departs from traditional matching methods in two important aspects. First, the synthetic control method does not impose a fixed number of matches for every treated unit. Second, instead of using a simple average of the matched units with equal weights, the synthetic control method matches each treated unit to a weighted average of untreated units with weights calculated to minimize the discrepancies between the treated unit and the synthetic control in the values of the matching variables. Synthetic control estimators retain, however, appealing properties of nearest neighbor matching estimators, in particular: sparsity and non-negativity of the weights, and weights that sum to one. Like for nearest neighbor matching estimators, most of the synthetic control weights are equal to zero and a small number of untreated units contribute positive weights to reproduce the counterfactual of each treated observation without the treatment. Sparsity and non-negativity of the weights, along with the fact that synthetic control weights sum to one and define a weighted average, are important features that allow incorporating expert knowledge to evaluate and interpret the estimated counterfactuals (see Abadie et al., 2015). As shown in Abadie et al. (2015), similar to the synthetic control estimator, a regression-based estimator of the counterfactual of interest – i.e., the outcome for the treated in the absence of an intervention – implicitly uses a linear combination of outcomes for the untreated with weights that sum to one. However, unlike synthetic control weights, regression weights are not explicit in the outcome in the procedure, they are not sparse, and they can be negative or greater than one, allowing unchecked extrapolation outside the support of the data and complicating the interpretation of the estimate and the nature of the implicit comparison. While most applications of the synthetic control framework have focused on cases where only one or a few aggregate units are exposed to the intervention of interest, the method has found recent applications in contexts with disaggregated data, where samples contain large numbers of treated and untreated units, and the interest lies on the average effect of the treatment among the treated (see, e.g., Acemoglu et al., 2016; Gobillon and Magnac, 2016; Kreif et al., 2016). In such settings, one could simply construct a synthetic control for an aggregate of all treated units. However, interpolation

biases may be much smaller if the estimator of the aggregate outcome that would have been observed for the treated in the absence of the treatment is based on the aggregation of multiple synthetic controls, one for each treated unit.

Using synthetic controls to estimate treatment effects with disaggregated data creates some practical challenges. In particular, when the values of the matching variables for a treated unit fall in the convex hull of the corresponding values for the donor pool, it may be possible to find multiple convex combinations of untreated units that perfectly reproduce the values of the matching variables for the treated observation. That is, the best synthetic control may not be unique. One practical consequence of the curse of dimensionality is that each particular treated unit is unlikely to fall in the convex hull of the untreated units, especially if the number of untreated units is small. As a result, lack of uniqueness is not often a problem in settings with one or a small number of treated units and, if it arises, it can typically be solved by ad-hoc methods, like increasing the number of covariates or by restricting the donor pool to units that are similar to the treated units. In settings with many treated and many untreated units, non-uniqueness may be an important consideration and a problem which is harder to solve.

More generally, in contrast to common aggregate data settings with a small donor pool (see, e.g., Abadie and Gardeazabal, 2003; Abadie et al., 2010), when there is a large number of units in the donor pool, single untreated units may provide close matches to the treated units in the sample. Therefore, in such a setting, the researcher faces a trade-off between minimizing the covariate discrepancy between each treated unit and its synthetic control as a whole (pure synthetic control case) and minimizing the covariate discrepancy between each treated unit and each unit that contributes to its synthetic control (pure matching case).

This article provides a generalized synthetic control framework for estimation and inference. We introduce a penalization parameter that trades off pairwise matching discrepancies with respect to the characteristics of each unit in the synthetic control against matching discrepancies with respect to the characteristics of the synthetic control unit as a whole. This type of penalization is aimed to reduce interpolation biases by prioritizing inclusion in the synthetic control of units that are close to the treated in the space of matching variables. Moreover, we show that as long as the penalization parameter is positive, the generalized synthetic control estimator is unique and sparse. If the value of the penalization parameter is close to zero, our procedure selects the synthetic control that minimizes the sum of pairwise matching discrepancies (among the synthetic controls that best reproduce the characteristic of the treated units). If the value of the penalization parameter is large, our estimator coincides with the pair-matching estimator. We study the formal properties of the penalized synthetic control estimator and propose data driven choices of the penalization parameter.

Our approach is in the spirit of using machine learning techniques to improve the synthetic control estimator or, more generally, to provide new tools for program evaluation problems. Following Doudchenko and Imbens (2016) that represent synthetic controls as a solution to complete an outcome matrix with missing entries, Athey et al. (2017) assumes an underlying sparse factor structure for the outcome under no treatment and

adapts matrix completion techniques to estimate a counterfactual. Their estimator penalizes the complexity of the factor structure, while our approach penalizes the discrepancy between the treated unit and each control unit that enters the synthetic unit. Amjad et al. (2018) study a ℓ_q-penalized version of the synthetic control after de-noising the outcome for the donor pool. The bias-correction that we propose has also been independently studied *e.g.* Ben-Michael et al. (2019), Arkhangelsky et al. (2018).

Section 2 presents the penalized synthetic control estimator and discusses several of its geometric properties. Section 3 studies its large sample properties. Section 4 discusses permutation inference. Section 5 presents ways to choose the penalization term. Sections 6 and 7 illustrate the properties of the estimator through simulations and empirical applications, respectively. Section 8 contains a summary of the article and conclusions. The appendix gathers the proofs.

2. Penalized Synthetic Control

2.1. Synthetic Control for Disaggregated Data

We code treatment using a binary variable, D, so $D = 1$ for treated individuals and $D = 0$ otherwise. To define treatment effects we adopt the potential outcome notation in Rubin (1974). Let Y_1 and Y_0 be random variables representing potential outcomes under treatment and under no treatment, respectively. The effect of the treatment is $Y_1 - Y_0$. Realized outcomes are defined as

$$Y = \begin{cases} Y_1 & \text{if } D = 1, \\ Y_0 & \text{if } D = 0. \end{cases}$$

Let X be a $(p \times 1)$-vector of pre-treatment predictors of Y_0. Consider the distributions of the triple (Y_1, Y_0, X) under treatment and no treatment, with $E[\cdot|D = 1]$ and $E[\cdot|D = 0]$ denoting the corresponding expectation operators, and $E[\cdot|X, D = 1]$ and $E[\cdot|X, D = 0]$ denoting expectations conditional on X. Let P_1 and P_0 be the probability measures that describe the distribution of X for treated and nontreated, respectively.

In contrast to Abadie et al. (2010, 2015) who focus on the case of one or a small number of treated units, we adopt a framework where units are sampled at random from some population of interest.

Assumption 3.1 (Sampling) $\{(Y_{1i}, X_i)\}_{i=1,\dots,n_1}$ *are* n_1 *independent draws from the distribution of* (Y_1, X) *and* $\{(Y_{0i}, X_i)\}_{i=n_1+1,\dots,n}$ *are* n_0 *independent draws from the distribution of* (Y_0, X).

Combining data for treated and nontreated we obtain the pooled sample, $\{(Y_i, D_i, X_i)\}_{i=1}^n$, $n = n_0 + n_1$. To simplify notation, we reorder the observations in the sample so that the n_1 treated observations are first and the n_0 untreated observations are last. The quantity of interest is the average treatment effect on the treated (ATET):

$$\tau = E[Y_1 - Y_0|D = 1]. \tag{3.1}$$

Assumption 3.2 (Nested support) *$P_1 \ll P_0$, that is, P_1 is absolutely continuous with respect to P_0.*

Assumption 3.3 (Unconfoundedness I) $E[Y_0|X, D = 1] = E[Y_0|X, D = 0]$.

Versions of Assumptions 3.2 and 3.3 are ubiquitous in the program evaluation literature (see, e.g., Imbens, 2004). Assumption 3.2 states that there is no value of X for which individuals are always treated. In other words, for any treated unit, it should be possible to find a non-treated unit with the same value of the covariates in the population. Assumption 3.3 states that conditionally on a set of observed covariates or confounding factors, X, the expected potential outcome without the treatment is the same for treated and control individuals. Graphical causal structures that support Assumption 3.3 are studied in Pearl (2000) and the subsequent literature.

Notice that, under these two assumptions, the counterfactual $E[Y_0|D = 1]$ can be expressed as a weighted average of the outcome among the untreated, yielding

$$\tau = E[Y|D = 1] - E[VY|D = 0], \tag{3.2}$$

where $V = dP_1/dP_0$. Many econometric estimators of τ based on Assumptions 3.2 and 3.3, whether explicitly or implicitly, employ a sample analog of equation (3.2),

$$\frac{1}{n_1}\sum_{i=1}^{n} Y_i D_i - \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1 - D_i)V_i. \tag{3.3}$$

Popular estimators of this type in micro-econometrics include most notably regression (Angrist and Pischke, 2009; Abadie et al., 2015), propensity score weighting (Rosenbaum and Rubin, 1983; Hirano et al., 2003) and matching (Smith and Todd, 2005). For example, in the case of the pair-matching estimator, the weight V_i given to control unit i is equal to an integer counting the number of times control unit i is the nearest neighbor of a treated unit, rescaled by n_0/n_1. The synthetic control method (Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015; Doudchenko and Imbens, 2016) also belongs to this class of estimators. It matches each treated unit to a "synthetic control", that is, a weighted average of untreated units with weights chosen to make the values of the predictors of the outcome variable of each synthetic control closely match the values of the same predictors for the corresponding treated units.

While these assumptions are enough to recover the average treatment effect in equation (3.1), identification of a wide variety of parameters can be attained by strengthening the identifying conditions as in Assumptions 3.2′ and 3.3′ below.

Assumption 3.2′ (Common support) *$P_1 \ll P_0$ and $P_0 \ll P_1$.*

Assumption 3.3′ (Unconfoundedness II) $Y_1, Y_0 \perp\!\!\!\perp D|X$.

Parameters identified by the addition of Assumptions 3.2′ and 3.3′ include quantile treatment effects, that is, differences in the quantiles of the distributions of potential outcomes

(Firpo, 2007), bounds on the distribution of the treatment effect (Firpo and Ridder, 2008), or counterfactual distributions (Chernozhukov et al., 2013), among others. They also include parameters describing conditional features of the distribution of potential outcomes (see, e.g., Crump et al., 2008) and regression parameters obtained after imposing the same distribution of X for treated and non-treated (Ho et al., 2007; Abadie and Spiess, 2016). While, for the sake of clarity, this article focuses on the estimation of average treatment effects, the generalized synthetic control method outlined here can be applied to estimate any of the parameters above. Moreover, Assumptions 3.1-3.3 are not the only possible identification conditions in a synthetic control setting, nor necessarily the least restrictive ones. In particular, Abadie et al. (2010) show that under a factor-structure condition on the regression residual of the outcome on the covariates for the untreated, using synthetic controls that match pre-treatment outcomes for the treated help control for unobserved confounding that arises from heterogeneity in the factor loadings.

For any $(p \times 1)$ real vector X and any $(p \times p)$ real symmetric positive-definite matrix Γ, define the norm $\|X\| = (X'\Gamma X)^{1/2}$. Because Γ is diagonalizable with strictly positive eigenvalues, we can always transform the vector X so that the matrix Γ becomes the $(p \times p)$ identity matrix. As a result, without loss of generality, we will consider only $\Gamma = I$. In the synthetic control framework, model selection – that is, the choice of the variables included in X – is operationalized through the choice Γ, which rescales or weights each predictor in X according to its predictive power on the outcome (see Abadie et al., 2010). In a setting with many treated and untreated units, the standard synthetic control estimation procedure is as follows:

1. For each treated unit, $i = 1, \ldots, n_1$, compute the n_0-vector of weights $W_i^* = (W_{i,n_1+1}^*, \ldots, W_{i,n}^*)$ that solves

$$\begin{aligned} \min_{W_i \in \mathbb{R}^{n_0}} & \left\| X_i - \sum_{j=n_1+1}^{n} W_{i,j} X_j \right\|^2 \\ \text{s.t.} \quad & W_{i,n_1+1} \geq 0, \ldots, W_{i,n} \geq 0, \\ & \sum_{j=n_1+1}^{n} W_{i,j} = 1, \end{aligned} \tag{3.4}$$

   where $W_{i,j}^*$ is the weight given to control unit j in the synthetic control unit corresponding to treated unit i.

2. Estimate τ using the mean difference between the realized outcome and the synthetic outcome for the treated

$$\widehat{\tau} = \frac{1}{n_1} \sum_{i=1}^{n_1} \left[Y_i - \sum_{j=n_1+1}^{n} W_{i,j}^* Y_j \right], \tag{3.5}$$

Notice that $\widehat{\tau}$ is the estimator in equation (3.3) reweighting each nontreated unit, $j = n_1 + 1, \ldots, n$, by $V_j = (n_0/n_1) \sum_{i=1}^{n_1} W_{i,j}^*$, with $W_{i,j}^* = 0$ for $i \geq n_1 + 1$.

While, to simplify notation, we described here a cross-sectional setting only, the extension to the more common panel data setting for synthetic controls is immediate and we will use it later (see Section 5).

2.2. Penalized Synthetic Control

The main contribution of this article is to propose a penalized version of the synthetic control estimator in equation (3.4). For treated unit i and given a positive penalization constant λ, the penalized synthetic control weights, $W^*_{i,j}(\lambda)$, solve

$$\begin{aligned} \min_{W_i \in \mathbb{R}^{n_0}} & \left\| X_i - \sum_{j=n_1+1}^{n} W_{i,j} X_j \right\|^2 + \lambda \sum_{j=n_1+1}^{n} W_{i,j} \|X_i - X_j\|^2 \\ \text{s.t.} \quad & W_{i,n_1+1} \geq 0, \ldots, W_{i,n} \geq 0, \\ & \sum_{j=n_1+1}^{n} W_{i,j} = 1. \end{aligned} \tag{3.6}$$

The penalized synthetic control estimator is then given by

$$\widehat{\tau}(\lambda) = \frac{1}{n_1} \sum_{i=1}^{n_1} \left[Y_i - \sum_{j=n_1+1}^{n} W^*_{i,j}(\lambda) Y_j \right]. \tag{3.7}$$

The tuning parameter λ sets the trade-off between componentwise and aggregate fit. The choice of the value of λ is important and will be discussed in Section 5. The penalized synthetic control estimator encompasses both the synthetic control estimator and the nearest-neighbor matching as special polar cases. At one end of the spectrum, as $\lambda \to 0$, the penalized estimator becomes the synthetic control that minimizes the sum of pairwise matching discrepancies among the set of synthetic controls that best reproduce the characteristics of the treated units. Our motivation to choose among synthetic controls that fit the treated unit equally well by minimizing the sum of pairwise matching discrepancies is to reduce worst-case interpolation biases. At the other end of the spectrum, as $\lambda \to \infty$, the penalized estimator becomes the one-match nearest-neighbor matching with replacement estimator in Abadie and Imbens (2006).

Let X_0 be the $(p \times n_0)$ matrix with column j equal to X_{n_1+j}, and let Δ_i be the $(n_0 \times 1)$ vector with j-th element equal to $\|X_i - X_{n_1+j}\|^2$. Moreover, let $\Delta_i^{NN} = \min_{j=1,\ldots,n_0} \|X_i - X_{n_1+j}\|^2$ be the smallest discrepancy between unit i and the units in the donor pool. Finally, let $W_i^*(\lambda)$ be a solution to (3.6), and $\Delta_i^*(\lambda) = \|X_i - X_0 W_i^*(\lambda)\|^2$ be the square of the discrepancy between unit i and the (penalized) synthetic control.

Lemma 3.1 (Discrepancy Bounds) *For any* $\lambda \geq 0$

$$0 \leq \Delta_i^*(\lambda) \leq \Delta_i^{NN},$$

and for $\lambda > 0$

$$\Delta_i^{NN} \leq \Delta_i' W_i^*(\lambda) \leq \frac{1+\lambda}{\lambda} \Delta_i^{NN}.$$

All proofs are in the appendix.

The first result in Lemma 3.1 states that the synthetic unit is contained in a closed ball of center X_i and radius equal to the distance to the nearest-neighbor, $\sqrt{\Delta_i^{NN}}$. The second result implies that the tuning parameter λ controls the compound discrepancy between the treated unit and the units that contribute to the synthetic control, $\Delta_i' W_i^*(\lambda)$.

Some remarks are in order to justify the choice of the penalization term in equation (3.6). First, notice that the penalty term is linear rather than quadratic in the weights. This has the advantage of producing easy-to-interpret sparse solutions, similarly to a matching procedure.

In addition, the optimization problem in (3.6) can be solved via quadratic programming, like the standard synthetic control in (3.5). To see why notice that we can express the optimization problem in (3.6) as

$$\begin{aligned} &\min_{W \in \mathbb{R}^{n_0}} (X_i - X_0 W)' (X_i - X_0 W) + \lambda \Delta_i' W \\ &\text{s.t.} \quad 1_{n_0}' W = 1, W \geq 0, \end{aligned} \tag{3.8}$$

where 1_{n_0} is the $(n_0 \times 1)$ vector of ones and the inequality restriction applies to each component of W.

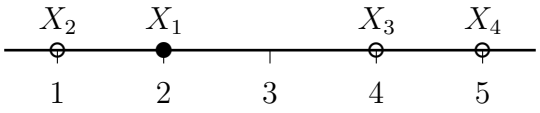
A third remark has to do with uniqueness of the solution. In the absence of the penalty term (that is, when $\lambda = 0$), the problem in (3.6) and (3.8) can be solved by projecting X_i on the convex hull of X_0. Existence of sparse solutions follows from Carathéodory's theorem. However, if $\lambda = 0$ the solution to the problem in (3.6) and (3.8) may not be unique if X_i belongs to the convex hull of the columns of X_0. Adopting $\lambda > 0$ penalizes solutions with potentially large interpolation biases created by large matching discrepancies and produces uniqueness and sparsity as stated in the following result.

Theorem 3.1 (Uniqueness and Sparsity) *Suppose that any submatrix composed by rows of* $[X_0' \; 1_{n_0} \; \Delta_i]$ *has full rank. Then, if* $\lambda > 0$ *the optimization problem in equation (3.6) admits a unique solution* $W_i^*(\lambda)$ *with at most* $p+1$ *non-zero components.*

Notice that the condition that any submatrix composed by rows of $[X_0' \; 1_{n_0} \; \Delta_i]$ has full rank implies that there are no two control units with the same values of the predictors. It also implies that there is no set of control units of cardinality $p+2$ or larger such that the values of the predictors belong to a sphere with center at X_i.

Example: Consider a simple numerical example with only one covariate. Suppose, there is one treated unit with $X_1 = 2$ and three control units with $X_2 = 1$, $X_3 = 4$ and $X_4 = 5$. This simple setting is depicted in Figure 3.1.
Notice that X_1 belongs to $[1, 5]$, the convex hull of the columns of X_0 and $\Delta_1 = (1, 4, 9)'$. Consider first the case with $\lambda = 0$. Then, $W^*(0) = (2/3, 1/3, 0)'$ and $W^{**}(0) = (3/4, 0, 1/4)'$ are the only two sparse solutions (with number of non-zero weights not

Figure 3.1: A simple example



greater than $p + 1 = 2$) to (3.6). The first sparse solution, $W^*(0)$, interpolates $X_1 = 2$ using $X_2 = 1$ and $X_3 = 4$. The second sparse solution, $W^{**}(0)$ is of lower quality relative to $W^*(0)$ in terms of compound discrepancy, as it uses an interpolation scheme that replaces X_3 with X_4, an observation farther away from X_1. As a result, $W^*(0)$ is preferred over $W^{**}(0)$ in terms of worst case interpolation bias (e.g., under a Lipschitz bound on $E[Y|X, D = 0]$). However, the better compound fit of $W^*(0)$ is not reflected in a better value in the objective function in (3.4). Moreover, because any convex combination of $W^*(0)$ and $W^{**}(0)$ is also a solution, the problem in (3.4) has an infinite number of solutions, $\mathcal{W}_0^* = \{aW^*(0) + (1-a)W^{**}(0) : a \in [0,1]\}$. Let $\bar{V}(a) = aW^*(0) + (1-a)W^{**}(0)$. The compound discrepancy of $\bar{V}(a)$ is

$$\Delta_i' \bar{V}(a) = 3 - a.$$

From Figure 1, it is apparent that $W^*(0)$, which is obtained making $a = 1$, produces the lowest compound discrepancy among all the solutions to equation (3.4).

When $\lambda > 0$, however, the program (3.6) has a unique solution, which is sparse:

$$W^*(\lambda) = \begin{cases} (2 + \lambda/2,\ 1 - \lambda/2,\ 0)'/3 & \text{if } 0 < \lambda \leq 2, \\ (1,\ 0,\ 0)' & \text{if } \lambda > 2. \end{cases}$$

Notice that $W^*(\lambda)$ never puts any weight on X_4. As $\lambda \to \infty$, $W^*(\lambda)$ selects the nearest-neighbor match, and as $\lambda \to 0$, $W^*(\lambda)$ converges to $W^*(0)$, the (non-penalized) synthetic control in $\mathcal{W}_0^*$ with the smallest compound discrepancy. □

The next theorem provides a characterization of the units contributing to a particular synthetic control, $X_0 W_i^*(\lambda)$ with $\lambda > 0$, as vertices of the face of the Delaunay complex (*i.e.* of a simplex) containing $X_0 W_i^*(\lambda)$ in the Delaunay tessellation of $X_{n_1+1}, \ldots, X_n$. For general references on Delaunay tessellations and related concepts, see Okabe et al. (2000); Boissonnat and Yvinec (1998).

Theorem 3.2 (Delaunay Property I) *Let $W_i^*(\lambda)$ be a solution to the penalized synthetic control problem in* (3.6) *with $\lambda > 0$. Consider the Delaunay tessellation induced by the columns of X_0. Then, for any control unit $j = n_1 + 1, \ldots, n$, such that X_j is not a vertex of the face of the Delaunay complex containing $X_0 W_i^*(\lambda)$, it holds that $W_{i,j}^*(\lambda) = 0$.*

This result along with the first part of Lemma 3.1, which bounds $\|X_i - X_0 W_i^*(\lambda)\|$, provides a notion of proximity between each treated unit X_i and the untreated units that contribute to its synthetic control. Theorem 3.2 provides also a simple way to compute the solution for the "pure synthetic control" case ($\lambda \to 0$) that does not entail the choice of an arbitrarily small value of λ to use in (3.6). Recall that when $\lambda = 0$, the

problem of minimizing $\|X_i - X_0 W\|$ subject to the weight constraints may have multiple (infinite) solutions, in which case $X_i = X_0 W$ for all solutions. In the presence of multiple solutions, the pure synthetic control case selects the solution that produces the lowest compound discrepancy, $W'\Delta_i$, among all W such that $X_i = X_0 W$. Directly solving (3.6) for $\lambda \to 0$ requires, in practice, a choice of a small value for λ. It also creates computational difficulties, as the minimization problem is close to one with multiple solutions and the dimension of W may be large. Theorem 3.2 provides a solution to these problems, because it implies that the solution of (3.6) for $\lambda \to 0$ can assign positive weights only to the vertices of the face in the Delaunay tessellation of $X_{n_1+1}, \ldots, X_n$ that contains the projection of X_i on the convex hull of the columns of X_0. As a result, it is enough to solve (3.6) allowing only positive weights on the observations that represent the vertices of the Delaunay face that contains the projection of of X_i on the convex hull of the columns of X_0. In high dimensional settings, however, the large computation costs of Delaunay triangulations may make this approach unfeasible.

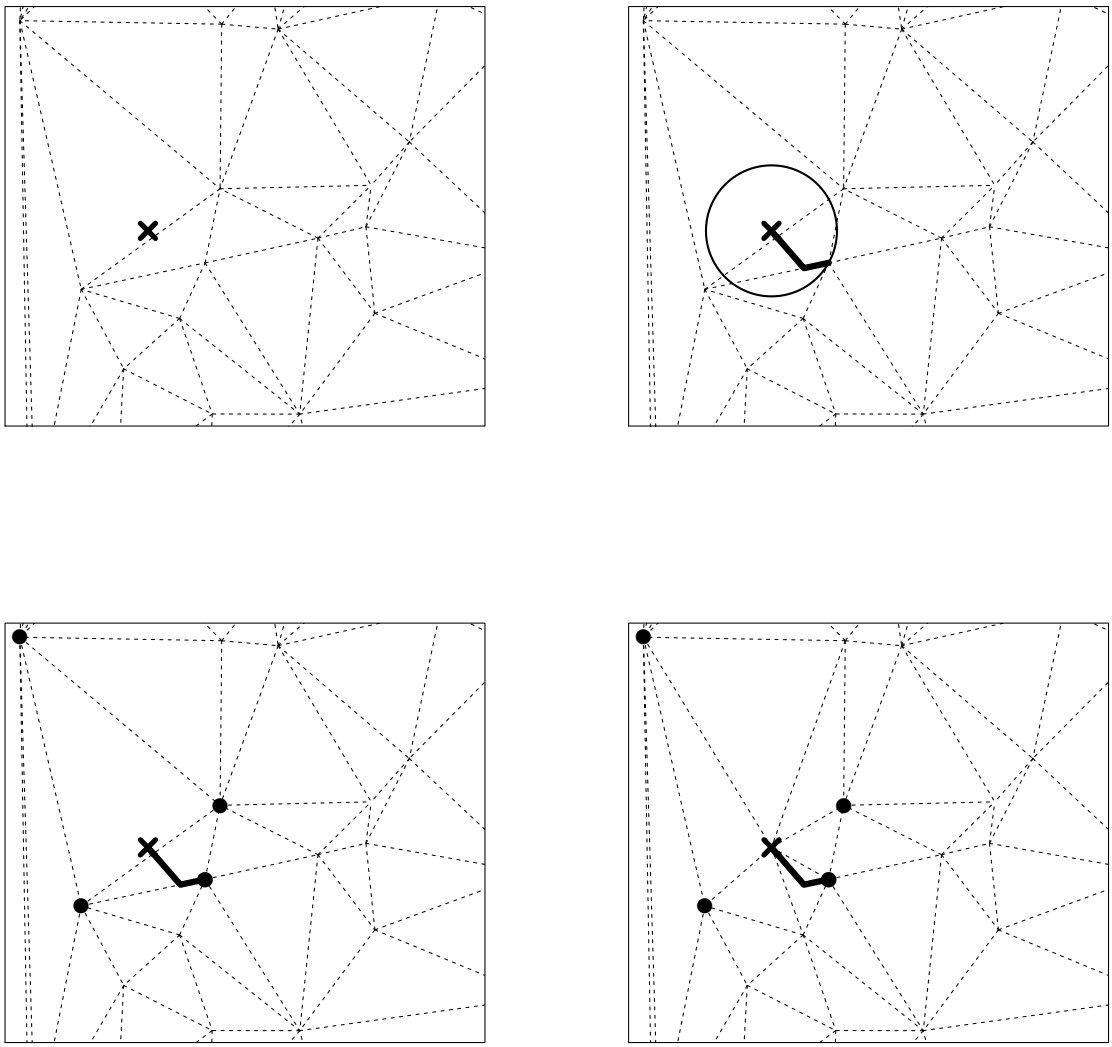
The next theorem shows that untreated units contributing to the synthetic control are connected to the treated unit through an edge in the augmented Delaunay tessellation induced by X_i and the columns of X_0.

Theorem 3.3 (Delaunay Property II) *Suppose that the columns of* $[X_i \; X_0]$ *are in General Quadratic Position. Let* $W_i^*(\lambda)$ *be a solution to the penalized synthetic control problem in* (3.6) *with* $\lambda > 0$. *Consider the Delaunay tessellation induced by the columns of* X_0 *and the treated* X_i, *and denote* $\mathcal{I}_i$ *the set of indices of points in* $\{X_{n_1+1}, \ldots, X_n\}$ *that are connected to* X_i *through a Delaunay edge. For any* $j \notin \mathcal{I}_i$, *it holds that* $W_{i,j}^*(\lambda) = 0$.

Theorem 3.3 provides a necessary condition for an untreated unit to take part in the synthetic unit: it has to be connected to the treated in the augmented tessellation. It therefore restricts the donor pool to these units connected to the treated and as such provides a way to decompose the computation of the synthetic control. Most importantly, this theorem also helps us controlling the probability of being assigned a non-zero weight, conditional on the donor pool, which we need to study the large-sample behavior of the synthetic control estimator (3.7). The assumption that the columns of $[X_i \; X_0]$ are in General Quadratic Position ensures the existence and the uniqueness of the augmented Delaunay tessellation (Okabe et al., 2000, Property D1). It requires the set of the $n_0 + 1$ points $\{X_i, X_{n_1+1}, \ldots, X_n\}$ to satisfy the following assumptions (Moller, 1994): (i) for $k = 2, \ldots, p$, no $k+1$ points lie in a $(k-1)$-dimensional hyperplane of $\mathbb{R}^p$ *(non-collinearity)* and (ii) when $n_0 + 1 \geq p + 2$, for $k \geq p + 2$, there does not exist a hypersphere such that k points are on this hypershpere and all other points are outside of it *(non-cosphericity)*. Notice that it is verified almost surely if the columns of $[X_i \; X_0]$ are distributed according to a measure absolutely continuous with respect to the Lebesgue measure.

Figure 3.2 illustrates Lemma 3.1 and Theorems 3.2-3.3. The top-left panel displays the treated (black cross) on the Delaunay triangulation of untreated units. The top-right panel draws the synthetic unit as λ changes (as λ increases, the solution drifts toward the nearest neighbor and away from the treated – solid black line) and the circle centered on the treated of radius equal to the distance between the treated and its nearest neighbor. Notice that the synthetic unit is never located outside of this circle, as per Lemma 3.1.

Figure 3.2: Geometric properties of penalized synthetic control estimator



The bottom left panel shows the four untreated (black dots) that have a non-zero weight across some solutions of the penalized synthetic control as λ changes. They are the vertices of the two triangles where the synthetic unit is located, as per Theorem 3.2. The bottom-right panel shows that these units are also connected to the treated in the augmented Delaunay triangulation, as per Theorem 3.3. However, notice that being connected to the treated is only a necessary condition and is not sufficient.

2.3. Bias-Corrected Synthetic Control

We will also consider bias-corrected versions of synthetic control estimator. We adopt a bias correction analogous to that implemented in Abadie and Imbens (2011) for matching estimators. Let $\mu_0(x) = E[Y|X = x, D = 0]$, and $\widehat{\mu}_0(x)$ be an estimator of $\mu_0(x)$. A

bias-corrected version of the synthetic control estimator in equation (3.7) is

$$\widehat{\tau}_{BC}(\lambda) = \frac{1}{n_1}\sum_{i=1}^{n_1}\left[\big(Y_i - \widehat{\mu}_0(X_i)\big) - \sum_{j=n_1+1}^{n} W^*_{i,j}(\lambda)\big(Y_j - \widehat{\mu}_0(X_j)\big)\right]. \tag{3.9}$$

In independent research, Ben-Michael et al. (2019) propose a related bias-correction for the synthetic control method.

3. Large Sample Properties

In this section we analyze the large sample properties of the penalized synthetic control estimator (3.7). Let $\mu_1(x) = E[Y|X = x, D = 1]$. And let $S_j(\lambda) = \sum_{i=1}^{n_1} W^*_{i,j}(\lambda)$ be the sum of weights given to untreated unit j across all synthetic units. Notice that $\widehat{\tau}(\lambda) - \tau = B_n(\lambda) + M_n(\lambda)$, where

$$B_n(\lambda) = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\mu_0(X_i) - \sum_{j=n_1+1}^{n} W^*_{i,j}(\lambda)\mu_0(X_j)\right),$$

$$M_n(\lambda) = \frac{1}{n_1}\sum_{i=1}^{n} D_i\left(\mu_1(X_i) - \mu_0(X_i) - \tau\right) + \frac{1}{n_1}\sum_{i=1}^{n}\left(D_i - (1-D_i)S_i(\lambda)\right)\left(Y_i - \mu_{D_i}(X_i)\right).$$

$B_n(\lambda)$ is a bias term, while $M_n(\lambda)$ has mean zero and can be expressed as a martingale with respect to a certain filtration. First, consider the bias term.

3.1. Bias

Assumption 3.4 (Regularity Conditions) *Suppose (i) P_0, the probability measure of X for the non-treated, admits a density that is bounded away from zero and above by a constant on compact and convex support $\mathcal{X}$; (ii) $\mu_0(.)$ is Lipschitz continuous on $\mathcal{X}$.*

Lemma 3.2 (Bias Bound) *Under Assumptions 3.1, 3.2, 3.3 and 3.4, if $\lambda > 0$:*

$$B_n(\lambda) = \mathcal{O}_P(n^{-1/p}).$$

The bias term of the estimator, $B_n(\lambda)$, exhibits a similar behavior to the one studied in Theorem 1 of Abadie and Imbens (2006) for matching estimators.

3.2. Consistency

For $d = 0, 1$, let $\sigma^2_d(x) = \text{var}\,(Y|X = x, D = d)$.

Assumption 3.5 (Regularity Conditions, Consistency) *Assume: (i) $E\big[|Y|\big|D = 1\big] < \infty$; (ii) $\sigma^2_0(.)$ is bounded on $\mathcal{X}$ by $\bar{\sigma}^2_0 < \infty$ a.s.; (iii) for $j = n_1 + 1, \ldots, n$, $(n_0/n_1^2)E[S_j(\lambda)^2|D = 0] \to 0$.*

Theorem 3.4 (Consistency) *Consider the estimator, $\widehat{\tau}(\lambda)$, defined in equation* (3.7) *with weights $W_{i,j}^{*}(\lambda)$ defined in equation* (3.6) *and $\lambda > 0$. Under Assumptions 3.1, 3.2, 3.3, 3.4 and 3.5:*

$$\widehat{\tau}(\lambda) - \tau = M_n(\lambda) + B_n(\lambda) \xrightarrow{p} 0, \text{ as } n_1, n_0 \to \infty.$$

3.3. Asymptotic Normality

In order to show asymptotic normality of the penalized synthetic control estimator, we employ the martingale representation derived in Abadie and Imbens (2012) for matching estimators. This representation allows us to use elegant results from martingale limit theory (Hall and Heyde, 1980) to circumvent the difficulty posed by the dependence between the sums of weights given to two different untreated units across all the synthetic controls. Notice that

$$\sqrt{n_1} M_n(\lambda) = \sum_{k=1}^{2n} \xi_{n,k},$$

where

$$\xi_{n,k} = \begin{cases} \dfrac{1}{\sqrt{n_1}} D_k \left(\mu_1(X_k) - \mu_0(X_k) - \tau\right) & \text{if } 1 \leq k \leq n, \\ \dfrac{1}{\sqrt{n_1}} \left(D_{k-n} - (1 - D_{k-n}) S_{k-n}(\lambda)\right) \left(Y_{k-n} - \mu_{D_{k-n}}(X_{k-n})\right) & \text{if } n+1 \leq k \leq 2n. \end{cases}$$

Let $\mathcal{X}_n = \{X_1, ..., X_n\}$, $\mathcal{D}_n = \{D_1, ..., D_n\}$ and consider $\mathcal{F}_{n,k} = \sigma(\mathcal{D}_n, X_1, ..., X_k)$ for $k \leq n$ and $\mathcal{F}_{n,k} = \sigma(\mathcal{D}_n, \mathcal{X}_n, Y_1, ..., Y_{k-n})$ for $k > n$. It follows that

$$\left\{ \sum_{k=1}^{j} \xi_{n,k}, \mathcal{F}_{n,j}, 1 \leq j \leq 2n \right\}$$

is a martingale.

Assumption 3.6 (Regularity Conditions, Asymptotic Normality) *Assume: (i) $E[Y^4 | X = x, D = d] < \infty$ and is uniformly bounded in x for $d = 0, 1$; (ii) $\sigma_d^2(.)$ for $d = 0, 1$ is bounded from above and away from zero; (iii) for $j = n_1 + 1, \ldots, n$, $(1/n_0) E[S_j(\lambda)^4 | D = 0] \to 0$; (iv) for $j = n_1 + 1, \ldots, n$, $k \neq j$, either $Cov(S_j(\lambda)^2 \sigma_0^2(X_j), S_k(\lambda)^2 \sigma_0^2(X_k)) \leq 0$, or $Cov(S_j(\lambda)^2 \sigma_0^2(X_j), S_k(\lambda)^2 \sigma_0^2(X_k)) \to 0$.*

Theorem 3.5 (Asymptotic Normality) *Consider the estimator $\widehat{\tau}(\lambda)$ defined in equation* (3.7) *with weights $W_{i,j}^{*}(\lambda)$ defined in equation* (3.6)*. Under Assumptions 3.1, 3.2, 3.3, 3.5 and 3.6, if $\lambda > 0$:*

$$\widehat{\sigma}^{-1}(\lambda) \sqrt{n_1} M_n(\lambda) = \widehat{\sigma}^{-1}(\lambda) \sqrt{n_1} \left(\widehat{\tau}(\lambda) - \tau - B_n(\lambda)\right) \xrightarrow{d} \mathcal{N}(0, 1), \text{ as } n_1, n_0 \to \infty,$$

where

$$\widehat{\sigma}^2(\lambda) := \frac{1}{n_1}\sum_{i=1}^{n} D_i\left(\mu_1(X_i) - \mu_0(X_i) - \tau\right)^2 + \frac{1}{n_1}\sum_{i=1}^{n} D_i\sigma_1^2(X_i) + (1-D_i)S_i(\lambda)^2\sigma_0^2(X_i).$$

In practice, as in Abadie and Imbens (2006, section 4), $\widehat{\sigma}^2(\lambda)$ can be replaced by an estimator that does not require consistent estimation of the unknown functions, $\mu_1()$, $\mu_0()$, $\sigma_1^2()$ and $\sigma_0^2()$.

Of independent interest, we develop methods in the proofs of Theorems 3.4 and 3.5 that can also be applied to derive the large sample distribution of any estimator of ATET of the form in equation (3.5), where the weights on the untreated depend on treatment status and covariates but not on outcomes, and they sum to one across the second index. This class includes the matching and hot-deck imputation estimators studied in Abadie and Imbens (2006, 2012).

3.4. Asymptotic Behavior of $S(\lambda)$

Assumptions 3.5(iii) and 3.6(iii)-(iv) regarding the asymptotic behavior of $S_j(\lambda)$ are high-level. This section provides low-level assumptions under which they hold.

Assumption 3.7 (Regularity Conditions on Covariate Distributions) *(i) P_0, the probability measure of X for the non-treated, admits a density that is Hölder continuous; (ii) P_1, the probability measure of X for the treated, admits a bounded density f_1 on $\mathcal{X}$, such that for $x \in \mathcal{X}$, $f_1(x) \leq \overline{f}_1$; (iii) there exist constants $\kappa < p$ and $C_{\partial\mathcal{X}} > 0$ such that the inside covering number of $\partial\mathcal{X}$, $N(\partial\mathcal{X}, \varepsilon) \leq C_{\partial\mathcal{X}}\varepsilon^{-\kappa}$; (iv) n_1/n_0 is bounded by a constant.*

The first three parts of Assumption 3.7 in addition to Assumption 3.4(i) are required to control the volume of the largest hypersphere that does not contain any untreated unit, using results from the stochastic geometry literature (Janson, 1987; Aaron et al., 2017). Intuitively, a control on the largest empty hypersphere is useful because a necessary condition for $W_{i,j}^*(\lambda) > 0$ is that X_i is linked to X_j in the Delaunay tesselation induced by X_i and the columns of X_0 (see Theorem 3.3). Such an event can only happen if X_i falls into the (empty) circumscribed hypersphere of any of the Delaunay simplex that has X_j as a vertex in the Delaunay tesselation induced by the columns of X_0. The volume of the largest empty hypersphere, *i.e.* of the largest hypersphere that does not contain any non-treated unit, provides a control on the conditional probability that $W_{i,j}^*(\lambda) > 0$ and ultimately on the behavior of $S_j(\lambda) = \sum_{i=1}^{n_1} W_{i,j}^*(\lambda)$. The last part of Assumption 3.7 requires that the number of treated is proportional to the number of untreated.

Lemma 3.3 (Control of $S(\lambda)$) *Under Assumptions 3.1, 3.2, 3.4(i) and 3.7, for any $j = n_1+1, \ldots, n$, and $\lambda > 0$, $S_j(\lambda) = \sum_{i=1}^{n_1} W_{i,j}^*(\lambda)$, the sum of weights given to control unit j across all synthetic units is such that for $m \geq 1$,*

$$\frac{n_0}{n_1^2} E\left[S_j(\lambda)^m | D = 0\right] \to 0, \text{ as } n_1, n_0 \to \infty.$$

Assumption 3.6(iv) is more complex to verify. For the sake of discussion, let us suppose that $\sigma_0^2(x)$ is constant over the support $\mathcal{X}$. Intuitively, the negative covariance of the squares could be viewed as a consequence of the Negative Association (NA) property of the random variables $S_{n_1+1}(\lambda), \ldots, S_n(\lambda)$. However, the NA property is not straightforward, as *e.g.* Theorem 2.6 in Joag-Dev and Proschan (1983) does not apply, even though for any partition (I, J) of $\{n_1 + 1, \ldots, n\}$, and any non-decreasing functions f_1, f_2, we have that $Cov\left(f_1\left(\sum_{j\in I} S_{n_1+j}(\lambda)\right), f_2\left(\sum_{j\in J} S_{n_1+j}(\lambda)\right)\right) \leq 0$ by Chebychev's inequality. Furthermore, Assumption 3.6(iv) would also hold – still by Chebychev's inequality – if $x \to E[S_{n_1+1}(\lambda)^2 | S_{n_1+2}(\lambda) = x]$ was a decreasing function, a result in the flavor of Efron (1965). Monte Carlo simulations using the design of Section 6 not reproduced in the present article provide evidence that Assumption 3.6(iv) indeed holds as we never reject it at a level of 5% for a wide range of possible values of n_1, n_0 and p.

4. Permutation Inference

The results of Section 3 provide an asymptotic approximation to the distribution of synthetic control estimators (as $n_1, n_0 \to \infty$). When large samples of treated and non-treated are available, those results provide the basis for inferential exercises on ATET. In other cases, however, the number of sample units may not be large enough to justify an asymptotic approximation to the distribution of the penalized synthetic control estimator.

In this section, we adapt the inferential framework in Abadie et al. (2010) to the penalized synthetic control estimators of Section 2. Like in Abadie et al. (2010), our inferential exercises compare the value of a test statistic to its permutation distribution induced by random reassignment of the treatment variable in the data set. This inferential exercise is exact by construction, regardless of the number of sample units. We next describe two possible implementations that employ different test statistics and permutation schemes. Alternative test statistics and permutation schemes are possible and, in practice, the choice among them should take into account the nature of the parameter(s) of interest (e.g., individual vs. aggregate effects), the characteristics of the intervention that is the object of the analysis and the structure of the data set. Randomized reassignment of the treatment in the data is taken here as a benchmark against which we evaluate the rareness of the sample value of a test statistic, and it may not reflect the actual and typically unknown treatment assignment process (see Abadie et al., 2010, 2015). Firpo and Possebom (2018) propose a procedure to assess the sensitivity of permutation inference to deviations from the reassignment benchmark.

4.1. Inference on Aggregate Effects

Here we outline a simple permutation procedure that employs a test statistic, $\widehat{T}$, that measures aggregate effects for the treated. Examples of aggregate statistics of this type are the synthetic controls estimators in equations (3.7) and (3.9). Similar to Abadie et al. (2010), in a panel data setting $\widehat{T}$ can be based on the ratio between the aggregate mean square prediction error in a post-intervention period $\mathcal{T}_1 \subseteq \{T_0 + 1, \ldots, T\}$ and a

pre-intervention period $\mathcal{T}_0 \subseteq \{1, \ldots, T_0\}$,

$$\sum_{t \in \mathcal{T}_1} \left(\sum_{i=1}^{n_1} \widehat{\tau}_{it}(\lambda) \right)^2 \Big/ \sum_{t \in \mathcal{T}_0} \left(\sum_{i=1}^{n_1} \widehat{\tau}_{it}(\lambda) \right)^2 . \tag{3.10}$$

Let $\boldsymbol{D}^{obs} = (D_1, ..., D_n)$ be the observed treatment assignment. We will write $\widehat{T}(\boldsymbol{D}^{obs})$ to indicate the value of the test statistic in the sample, and $\widehat{T}(\boldsymbol{D})$ to indicate the value of the test statistic when the treatment values are reassigned in the data as indicated in $\boldsymbol{D}$. The test is as follows:

1. Compute the treatment effect estimate in the original sample $\widehat{T}(\boldsymbol{D}^{obs})$.

2. At each iteration, $b = 1, \ldots, B$, permute at random the components of $\boldsymbol{D}^{obs}$ to obtain $\widehat{T}(\boldsymbol{D}^{(b)})$.

3. Calculate p-values as the frequency across iterations of values of $\widehat{T}(\boldsymbol{D}^{(b)})$ more extreme than $\widehat{T}(\boldsymbol{D}^{obs})$. Typically, for two-sided tests:

$$\widehat{p} = \frac{1}{B+1} \left(1 + \sum_{b=1}^{B} \mathbf{1} \left\{ |\widehat{T}(\boldsymbol{D}^{(b)})| \geq |\widehat{T}(\boldsymbol{D}^{obs})| \right\} \right).$$

   For one sided tests:

$$\widehat{p} = \frac{1}{B+1} \left(1 + \sum_{b=1}^{B} \mathbf{1} \left\{ \widehat{T}(\boldsymbol{D}^{(b)}) \geq \widehat{T}(\boldsymbol{D}^{obs}) \right\} \right),$$

   or

$$\widehat{p} = \frac{1}{B+1} \left(1 + \sum_{b=1}^{B} \mathbf{1} \left\{ \widehat{T}(\boldsymbol{D}^{(b)}) \leq \widehat{T}(\boldsymbol{D}^{obs}) \right\} \right).$$

4.2. Inference Based on the Sum of Rank Statistics of Unit-Level Treatment Effects Estimates

Similar to Dube and Zipperer (2015), we propose a test based on the rank statistics of the unit-level treatment effects. Unlike the test in Dube and Zipperer (2015), we calculate the permutation distribution directly from the data. The test we employ is based on the sum of ranks of individual treatment effects in the ordered sample combining the $n_1 \times (B+1)$ unit-level treatment effects for the actual assignments and B random permutations. Individual treatment effects, $\widehat{T}_i$, may be based on differences in outcomes between treated and synthetic controls,

$$Y_i - \sum_{j=n_1+1}^{n} W^*_{i,j}(\lambda) Y_j,$$

bias corrected versions of the unit-level treatment effects,

$$(Y_i - \widehat{\mu}_0(X_i)) - \left(\sum_{j=n_1+1}^{n} W^*_{i,j}(\lambda) Y_j - \widehat{\mu}_0(X_j) \right),$$

or unit-level versions of the mean squared prediction error ratio in equation (3.10). The test is implemented as follows:

1. Compute unit-level measures treatment effects for the treated, $\widehat{T}_i$, for $i = 1, \ldots, n_1$, under the actual treatment assignment, $\boldsymbol{D}^{obs}$.

2. At each iteration $b = 1, \ldots, B$, permute at random the components of $\boldsymbol{D}^{obs}$ to obtain treatment effects $\widehat{T}_i(\boldsymbol{D}^{(b)})$ for the treated. Denote these estimates $\widehat{T}_1^{(b)}, \ldots, \widehat{T}_{n_1}^{(b)}$ (in arbitrary order).

3. Calculate the ranks $R_1, \ldots, R_{n_1}$, $R_1^{(1)}, \ldots, R_{n_1}^{(1)}$, ..., $R_1^{(B)}, \ldots, R_{n_1}^{(B)}$ associated to the $n_1 \times (B+1)$ individual treatment effect estimates $\widehat{T}_1, \ldots, \widehat{T}_{n_1}$, $\widehat{T}_1^{(1)}, \ldots, \widehat{T}_{n_1}^{(1)}$, ..., $\widehat{T}_1^{(B)}, \ldots, \widehat{T}_{n_1}^{(B)}$ (or of their absolute values or negative values) and the sums of ranks for each permutation, $SR = \sum_{i=1}^{n_1} R_i$, $SR^{(b)} = \sum_{i=1}^{n_1} R_i^{(b)}$, $b = 1, \ldots, B$.

4. Calculate p-values as:

$$\widehat{p} = \frac{1}{B+1} \left(1 + \sum_{b=1}^{B} \mathbf{1}\left\{ SR^{(b)} \geq SR \right\} \right).$$

5. Penalty Choice

We present two data-driven selectors for the penalty term, λ. In the context of treatment effects estimation, cross-validation is complicated by the absence of data on a "ground truth" (that is, on the values of Y_0 for the treated units in the post-intervention periods, see Athey and Imbens, 2016). Since synthetic controls are often applied to panel data, we consider a balanced panel data setting with T periods and $T_0 < T$ pre-intervention periods. We define Y_{it} as the outcome for unit i at time t. Adaptation of (3.6) and (3.7) to the panel data setting is straightforward by allowing X_i to potentially include multiple pre-intervention values of the outcome variable and of other predictors of post-intervention outcomes.

The first selector proposed in this section is based on cross-validation on the outcomes on the untreated units in the post-intervention period. The second selector uses a strategy similar to the model selection procedure in Abadie et al. (2015), minimizing Mean Squared Prediction Error (MSPE) in a hold-out pre-intervention period.

5.1. Leave-One-Out Cross-Validation of Post-Intervention Outcomes for the Untreated

This section discusses a leave-one-out cross-validation procedure to select λ by minimizing mean squared prediction error for the untreated units in the post-intervention period. The procedure is as follows:

1. For each control unit $i = n_1 + 1, ..., n$, and each post-intervention period, $t = T_0 + 1, \ldots, T$, calculate

$$\widehat{\tau}_{it}(\lambda) = Y_{it} - \sum_{\substack{j=n_1+1 \\ j \neq i}}^{n} W^*_{i,j}(\lambda) Y_{jt},$$

   where $W^*_{i,j}(\lambda)$ is a synthetic control for unit i that is produced by the donor pool $\{n_1 + 1, \ldots, n\} \backslash \{i\}$.

2. Choose λ to minimize some measure of loss, such as the mean squared prediction error for the individual outcomes,

$$\frac{1}{n_0(T - T_0)} \sum_{i=n_1+1}^{n} \sum_{t=T_0+1}^{T} \left(\widehat{\tau}_{it}(\lambda) \right)^2.$$

5.2. Pre-Intervention Holdout Validation on the Outcomes of the Treated

An alternative selector of λ is based on validation over the outcomes for the treated on a hold out pre-intervention period. This is similar in spirit to the model selection procedure in Abadie et al. (2015). To simplify the exposition, and because it may be the most natural choice, we will only describe the case where the training and validation periods come immediately before the intervention, although other choices are possible. Let h and k be the lengths of the training and validation periods, respectively. The validation period comprises the k periods immediately before the intervention, and the training period comprises the h periods immediately before the validation period. The procedure is as follows:

1. For each treated individual, i, and validation period, $t \in \{T_0 - k + 1, \ldots, T_0\}$, compute

$$\widehat{\tau}_{it}(\lambda) = Y_{it} - \sum_{j=n_1+1}^{n} W^*_{i,j}(\lambda) Y_{jt},$$

   where $W^*_{i,j}$ solve (3.6) with $X_1, \ldots, X_n$ measured in the training period.

2. Choose λ to minimize a measure of error, such as the sum of the squared prediction

for the individual outcomes,

$$\sum_{i=1}^{n_1} \sum_{t=T_0-k+1}^{T_0} \left(\widehat{\tau}_{it}(\lambda)\right)^2,$$

or the squared prediction error of the aggregate outcomes,

$$\sum_{t=T_0-k+1}^{T_0} \left(\sum_{i=1}^{n_1} \widehat{\tau}_{it}(\lambda)\right)^2.$$

Notice that the cross-validation procedures delineated can also be applied here to guide model selection (i.e., choice of the weight for each covariate in the minimization program) as in Abadie et al. (2015).

6. Simulations

This section reports the results of a Monte Carlo experiment that investigates the finite sample properties of the penalized synthetic control estimator relative to its unpenalized version ($\lambda = 0$) and to the nearest-neighbor matching estimator in a panel data framework.

The data generating process is as follows. Let X_{mi} be the m-th component of X_i. The simulation design includes two periods: a pre-intervention period ($t = 1$), and a post-intervention period ($t = 2$). Irrespective of the treatment status, the outcome at date $t \in \{1, 2\}$ is generated by $Y_{it} = \left(\sum_{j=1}^{p} X_{mi}^r\right)/\beta + \varepsilon_{it}$ with r a positive real governing the degree of linearity of the outcome function. Hence, the treatment effect is zero. For any t, $\varepsilon_{it} \perp\!\!\!\perp X_i$ and ε_{it} is standard normal. For the n_1 treated units, X_i, is a vector of dimension p with i.i.d. entries uniformly distributed on $[.1, .9]$. For the n_0 control units, X_i is a vector of the same dimension with i.i.d. entries distributed as $\sqrt{U}$, where U is uniform on $[0, 1]$ We set $\beta = \sqrt{\text{var}\left(\sum_{j=1}^{p} X_{mi}^r | D_i = 1\right)}$, so that $\text{var}(Y_{i,t}|D_i = 1) = 2$ and the signal-to-noise ratio for the treated is equal to one.

We compare the performances of synthetic control and matching estimators. We will consider these two estimators with a fixed choice and a data-driven choice of λ and M. Under the fixed procedure, we impose $\lambda \to 0$ for the synthetic control and $M = 1$ in the matching estimator, encompassing both polar cases of the penalized synthetic control estimator highlighted in this paper. The case $\lambda \to 0$ is referred to as the "pure synthetic control". Among all the solutions to the unpenalized synthetic control optimization problem in equation (3.4), it selects the one with the smallest componentwise matching discrepancy, $\sum_{j=n_1+1}^{n} W_{i,j}\|X_i - X_j\|^2$. The computation of the pure synthetic control estimator is based on the result in Theorem 3.2 and discussion thereafter. The pure synthetic control estimator is not to be confused with the non-penalized synthetic control ($\lambda = 0$), for which we also report results, and which does not take into account the compound discrepancy. The data-driven choice of λ and M uses the first period outcome

to minimize the mean square error (MSE) over that period. In other words, we follow the second procedure in Section 5. At each simulation step, λ and M are chosen so as to minimize

$$MSE(\lambda) = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(Y_{i1} - \sum_{j=n_1+1}^{n} W^*_{i,j}(\lambda)Y_{j1}\right)^2,$$

and

$$MSE(M) = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(Y_{i1} - \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_{j1}\right)^2,$$

respectively, where $\mathcal{J}_M(i)$ is the set of indices of the M control units that are the nearest to treated unit i as measured by the euclidean norm. We also report a bias-corrected version of the estimators as in Section 2.3, based on a linear specification.

The degree of the outcome function, r, is the key parameter governing the relative performances of the candidate estimators. When $r = 1$, the outcome function is linear, in which case we expect the unpenalized and pure synthetic control estimators to do well, while the 1-to-1 matching should do relatively worse. In this setting, we expect the data driven value of λ be small. As r increases, the unpenalized and pure synthetic control estimators should suffer from a larger interpolation bias, while the performance of the 1-to-1 matching should improve. We expect the data-driven value of λ to increase with r.

For each configuration and each estimator, we report four statistics computed on the treated sample in the second period. The first is the individual-level MSE defined as

$$\frac{1}{B}\sum_{b=1}^{B}\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\widehat{\tau}^{(b)}_{i2}\right)^2.$$

The second is the aggregate-level MSE

$$\frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\widehat{\tau}^{(b)}_{i2}\right)^2.$$

The third is the aggregate absolute bias

$$\left|\frac{1}{B}\sum_{b=1}^{B}\frac{1}{n_1}\sum_{i=1}^{n_1}\widehat{\tau}^{(b)}_{i2}\right|.$$

The last is the average sparsity defined as the average number of control units used in the match to a given treated unit, i.e., number of non-zero entries of $W^*_i(\lambda)$ or number of matches in the optimized matching procedure.

The results are reported in Tables 3.1, 3.2 and 3.3 for $n_0 \in \{20, 40, 100\}$ respectively, each

time with $n_1 = 10$. Table 3.4 reports results for $n_1 = 100$, $n_0 = 500$. For Table 3.4, the pure synthetic control is omitted because of the high computational cost of calculating Delaunay triangulations for the setting in the table. Each table is divided into sixteen blocks corresponding to a particular value of (p, r). Each block is divided into two parts: the upper half reports the results without bias-correction and the lower half reports results with a bias-correction based on a linear specification of the regression function. Results are color-coded column-by-column within each half-block on a continuous color scale. For the upper half-block, the scale varies from dark blue (minimum column value) to light yellow (maximum column value). For the lower half-block, the scale varies from bright red (minimum column value) to light yellow (maximum column value).

Looking at Tables 3.1-3.3, several observations can be made. First, both the penalized and the pure synthetic control estimators consistently outperform the matching procedures across all three measures of performance. This advantage appears to be increasing with p, the dimension of the covariates. Second, the unpenalized synthetic control estimator shows mixed results, especially when p is small, but catches up with the pure synthetic control estimator as p increases, which is expected. Indeed, the pure and unpenalized synthetic control estimators coincide for treated units outside of the convex hull of the untreated. And as the dimensionality of the matching variables increases, the probability that a treated unit falls outside the convex hull of the untreated becomes large. Third, the advantage of the penalized and pure synthetic control estimators with respect to the bias slightly decreases as the degree of the outcome function r increases. When r is relatively large, the matching procedure displays a low bias as expected, albeit at the expense of a very large individual MSE. These three observations are magnified in Table 3.4 where the penalized synthetic control performs consistently well in each of the sixteen blocks. The biases of the estimators go down substantially when we adopt the bias-correction procedure of Section 2.3. Here, it is more difficult to rank estimators based on the simulation as the amount of bias corrected by the procedure is different for each estimator in a way that may be directly linked to the simulation design. That said, the overall patterns of relative performance of the bias-corrected estimators is similar to that of the the non-corrected estimators, albeit with more muted differences in performance.

Overall, Tables 3.1-3.4 give evidence that the penalized synthetic control estimator strikes a favorable bias-variance trade-off by combining the strength of matching and (unpenalized) synthetic control.

7. Empirical Applications

7.1. The Value of Connections in Turbulent Times, Acemoglu et al. (2016)

We revisit Acemoglu et al. (2016) which analyzes the effect of the announcement of the appointment of Tim Geithner as Treasury Secretary on November 21, 2008 on stock returns of firms that were connected to him. To choose λ we employ the pre-intervention holdout procedure of Section 5.2. The training sample uses stock returns over a 250-day window that ends 30 days prior to the Geithner announcement. The validation sample to select the tuning parameter λ uses returns on the following 30-day window. Abnormal

Table 3.1: Monte-Carlo Simulations, $n_1 = 10$, $n_0 = 20$

	$r = 1$				$r = 1.2$				$r = 1.4$				$r = 2$			
	MSE indiv.	MSE mean	|Bias|	Sparsity	MSE indiv.	MSE mean	|Bias|	Sparsity	MSE indiv.	MSE mean	|Bias|	Sparsity	MSE indiv.	MSE mean	|Bias|	Sparsity
$p = 2$, average number of treated outside convex hull: 4.35																
Pen. Synth.	1.3681	0.6014	0.2096	2.2814	1.3618	0.5987	0.2096	2.2823	1.3567	0.5964	0.2108	2.2853	1.3493	0.5919	0.2135	2.2796
Unpen. Synth.	1.2953	0.6297	0.2084	11.7803	1.2921	0.6353	0.2367	11.7803	1.2914	0.6430	0.2646	11.7803	1.3023	0.6767	0.3466	11.7803
Pure Synth.	1.3437	0.6008	0.2009	2.5124	1.3395	0.6000	0.2062	2.5124	1.3364	0.5998	0.2112	2.5124	1.3319	0.6017	0.2260	2.5124
Matching	1.5280	0.6368	0.2357		1.5233	0.6330	0.2293		1.5198	0.6295	0.2229		1.5157	0.6219	0.2060	
Opt. Matching	1.3603	0.6749	0.4174	4.5260	1.3541	0.6709	0.4095	4.4750	1.3493	0.6673	0.4033	4.4100	1.3381	0.6591	0.3935	4.4120
Pen. Synth. (BC)	1.3261	0.5714	0.0045		1.3263	0.5713	0.0061		1.3281	0.5715	0.0154		1.3441	0.5773	0.0423	
Unpen. Synth. (BC)	1.2535	0.5968	0.0140		1.2554	0.5970	0.0320		1.2608	0.5995	0.0497		1.2935	0.6183	0.1030	
Pure Synth. (BC)	1.3034	0.5708	0.0065		1.3041	0.5700	0.0015		1.3068	0.5704	0.0036		1.3234	0.5766	0.0175	
Matching (BC)	1.4492	0.5948	0.0031		1.4504	0.5948	0.0148		1.4539	0.5962	0.0324		1.4744	0.6071	0.0814	
Opt. Matching (BC)	1.2271	0.5285	0.0063		1.2293	0.5292	0.0303		1.2359	0.5319	0.0531		1.2676	0.5505	0.1184	
$p = 4$, average number of treated outside convex hull: 8.84																
Pen. Synth.	1.4731	0.8102	0.5713	2.9922	1.4749	0.8215	0.5868	2.9868	1.4768	0.8344	0.6024	2.9741	1.4902	0.8659	0.6398	2.9361
Unpen. Synth.	1.4453	0.8016	0.5590	4.9819	1.4482	0.8199	0.5860	4.9819	1.4529	0.8387	0.6120	4.9819	1.4771	0.8983	0.6873	4.9819
Pure Synth.	1.4480	0.8012	0.5600	3.3532	1.4509	0.8178	0.5839	3.3532	1.4553	0.8347	0.6068	3.3532	1.4768	0.8877	0.6728	3.3532
Matching	1.7010	0.8992	0.6264		1.6989	0.8994	0.6240		1.6980	0.8996	0.6212		1.7033	0.9022	0.6133	
Opt. Matching	1.5795	0.9682	0.7764	3.6150	1.5787	0.9698	0.7750	3.5250	1.5787	0.9709	0.7712	3.4870	1.5763	0.9789	0.7711	3.4050
Pen. Synth. (BC)	1.3051	0.6096	0.0078		1.3096	0.6132	0.0197		1.3180	0.6204	0.0317		1.3601	0.6493	0.0715	
Unpen. Synth. (BC)	1.2897	0.6168	0.0151		1.2938	0.6211	0.0157		1.3024	0.6276	0.0167		1.3491	0.6574	0.0181	
Pure Synth. (BC)	1.2927	0.6156	0.0141		1.2969	0.6200	0.0178		1.3050	0.6266	0.0219		1.3488	0.6558	0.0326	
Matching (BC)	1.4455	0.6431	0.0045		1.4485	0.6468	0.0248		1.4556	0.6533	0.0540		1.4967	0.6859	0.1372	
Opt. Matching (BC)	1.2815	0.6003	0.0132		1.2878	0.6038	0.0445		1.2990	0.6121	0.0767		1.3469	0.6527	0.1671	
$p = 8$, average number of treated outside convex hull: 10.00																
Pen. Synth.	1.8514	1.3112	1.1722	3.7048	1.8707	1.3385	1.2007	3.6512	1.8874	1.3615	1.2250	3.5926	1.9428	1.4326	1.2958	3.4703
Unpen Synth.	1.8275	1.3144	1.1826	4.3612	1.8494	1.3498	1.2216	4.3612	1.8717	1.3843	1.2589	4.3612	1.9426	1.4846	1.3650	4.3612
Pure Synth.	1.8275	1.3144	1.1826	4.3601	1.8494	1.3498	1.2216	4.3601	1.8717	1.3843	1.2589	4.3601	1.9426	1.4846	1.3650	4.3601
Matching	2.0989	1.3945	1.2228		2.1094	1.4080	1.2361		2.1204	1.4206	1.2480		2.1580	1.4569	1.2802	
Opt. Matching	2.0103	1.4916	1.3637	3.2010	2.0191	1.5073	1.3789	3.0990	2.0299	1.5179	1.3874	2.9480	2.0575	1.5491	1.4146	2.7920
Pen. Synth. (BC)	1.5372	0.8046	0.0167		1.5440	0.8090	0.0405		1.5556	0.8176	0.0670		1.6139	0.8557	0.1408	
Unpen. Synth. (BC)	1.5172	0.7980	0.0117		1.5213	0.8007	0.0283		1.5303	0.8062	0.0451		1.5824	0.8363	0.0926	
Pure Synth. (BC)	1.5172	0.7980	0.0116		1.5213	0.8008	0.0282		1.5304	0.8062	0.0451		1.5824	0.8363	0.0926	
Matching (BC)	1.6450	0.8274	0.0310		1.6501	0.8320	0.0665		1.6604	0.8402	0.1021		1.7179	0.8838	0.2046	
Opt. Matching (BC)	1.5497	0.8108	0.0220		1.5548	0.8142	0.0564		1.5676	0.8247	0.0927		1.6244	0.8671	0.1958	
$p = 10$, average number of treated outside convex hull: 10.00																
Pen. Synth.	2.0497	1.5653	1.4361	3.9785	2.0759	1.6012	1.4736	3.9163	2.1005	1.6322	1.5060	3.8594	2.1721	1.7179	1.5894	3.6503
Unpen Synth.	2.0307	1.5705	1.4510	4.7351	2.0618	1.6144	1.4975	4.7351	2.0928	1.6568	1.5421	4.7351	2.1871	1.7789	1.6683	4.7351
Pure Synth.	2.0307	1.5705	1.4510	4.7351	2.0618	1.6144	1.4975	4.7351	2.0928	1.6568	1.5421	4.7351	2.1871	1.7789	1.6683	4.7351
Matching	2.2908	1.6316	1.4763		2.3075	1.6520	1.4956		2.3241	1.6710	1.5132		2.3760	1.7249	1.5610	
Opt. Matching	2.2098	1.7203	1.5950	3.0870	2.2270	1.7413	1.6142	2.9670	2.2438	1.7601	1.6319	2.8680	2.2917	1.8147	1.6816	2.7000
Pen. Synth. (BC)	1.7845	1.0655	0.0166		1.7881	1.0679	0.0427		1.7961	1.0721	0.0714		1.8489	1.1045	0.1600	
Unpen. Synth. (BC)	1.7705	1.0598	0.0172		1.7725	1.0604	0.0383		1.7794	1.0638	0.0596		1.8264	1.0888	0.1204	
Pure Synth. (BC)	1.7705	1.0598	0.0172		1.7725	1.0604	0.0383		1.7794	1.0638	0.0596		1.8264	1.0888	0.1204	
Matching (BC)	1.8667	1.0867	0.0245		1.8691	1.0881	0.0626		1.8769	1.0931	0.1008		1.9285	1.1272	0.2112	
Opt. Matching (BC)	1.7970	1.0698	0.0192		1.8008	1.0724	0.0568		1.8103	1.0779	0.0940		1.8629	1.1130	0.2027	

Table 3.2: Monte-Carlo Simulations, $n_1 = 10$, $n_0 = 40$

	$r = 1$				$r = 1.2$				$r = 1.4$				$r = 2$			
	MSE indiv.	MSE mean	|Bias|	Sparsity	MSE indiv.	MSE mean	|Bias|	Sparsity	MSE indiv.	MSE mean	|Bias|	Sparsity	MSE indiv.	MSE mean	|Bias|	Sparsity
$p = 2$, average number of treated outside convex hull: 2.88																
Pen. Synth.	1.2955	0.4948	0.1258	2.4458	1.2932	0.4934	0.1275	2.4432	1.2920	0.4945	0.1296	2.4466	1.2876	0.4940	0.1332	2.4393
Unpen. Synth.	1.1985	0.5144	0.0973	27.6213	1.1970	0.5211	0.1420	27.6213	1.1989	0.5322	0.1850	27.6213	1.2202	0.5835	0.3059	27.6213
Pure Synth.	1.2695	0.4923	0.1132	2.6839	1.2673	0.4927	0.1192	2.6839	1.2658	0.4933	0.1245	2.6839	1.2641	0.4963	0.1385	2.6839
Matching	1.4757	0.5442	0.1626		1.4715	0.5410	0.1574		1.4684	0.5382	0.1523		1.4636	0.5322	0.1389	
Opt. Matching	1.2582	0.5516	0.3226	5.8850	1.2498	0.5483	0.3216	5.9820	1.2440	0.5433	0.3167	6.0020	1.2333	0.5350	0.3034	6.1120
Pen. Synth. (BC)	1.2707	0.4749	0.0103		1.2726	0.4757	0.0067		1.2754	0.4785	0.0036		1.2852	0.4853	0.0081	
Unpen. Synth. (BC)	1.1772	0.5024	0.0083		1.1790	0.5035	0.0313		1.1847	0.5091	0.0693		1.2183	0.5449	0.1762	
Pure Synth. (BC)	1.2494	0.4782	0.0076		1.2504	0.4791	0.0085		1.2524	0.4805	0.0089		1.2623	0.4865	0.0088	
Matching (BC)	1.4237	0.5078	0.0224		1.4245	0.5077	0.0108		1.4266	0.5086	0.0006		1.4385	0.5148	0.0318	
Opt. Matching (BC)	1.1637	0.4380	0.0071		1.1612	0.4376	0.0126		1.1648	0.4388	0.0322		1.1893	0.4570	0.0915	
p = 4, average number of treated outside convex hull: 7.66																
Pen. Synth.	1.3772	0.6611	0.4131	3.3715	1.3769	0.6693	0.4291	3.3603	1.3786	0.6797	0.4459	3.3563	1.3897	0.7059	0.4848	3.3040
Unpen. Synth.	1.3373	0.6500	0.4034	10.9692	1.3397	0.6704	0.4395	10.9692	1.3446	0.6920	0.4744	10.9692	1.3725	0.7628	0.5745	10.9692
Pure Synth.	1.3470	0.6496	0.4024	3.7334	1.3489	0.6646	0.4293	3.7334	1.3524	0.6801	0.4550	3.7334	1.3712	0.7297	0.5282	3.7334
Matching	1.6240	0.7609	0.4792		1.6202	0.7572	0.4749		1.6177	0.7537	0.4704		1.6183	0.7462	0.4584	
Opt. Matching	1.4690	0.8333	0.6701	4.6600	1.4634	0.8299	0.6667	4.5690	1.4598	0.8283	0.6653	4.5340	1.4529	0.8283	0.6662	4.5050
Pen. Synth. (BC)	1.2462	0.5056	0.0031		1.2487	0.5058	0.0056		1.2527	0.5061	0.0066		1.2815	0.5177	0.0201	
Unpen. Synth. (BC)	1.2107	0.5047	0.0018		1.2128	0.5056	0.0201		1.2189	0.5088	0.0377		1.2565	0.5296	0.0887	
Pure Synth. (BC)	1.2214	0.5030	0.0008		1.2229	0.5032	0.0099		1.2274	0.5049	0.0183		1.2551	0.5166	0.0423	
Matching (BC)	1.4267	0.5576	0.0096		1.4283	0.5573	0.0338		1.4330	0.5590	0.0579		1.4615	0.5739	0.1254	
Opt. Matching (BC)	1.2074	0.4860	0.0061		1.2085	0.4868	0.0210		1.2132	0.4887	0.0480		1.2428	0.5106	0.1209	
p = 8, average number of treated outside convex hull: 9.97																
Pen. Synth.	1.6953	1.1269	0.9992	4.3286	1.7109	1.1534	1.0274	4.2726	1.7261	1.1782	1.0546	4.2336	1.7735	1.2441	1.1228	4.0561
Unpen Synth.	1.6654	1.1153	0.9917	5.0143	1.6846	1.1511	1.0323	5.0143	1.7046	1.1860	1.0713	5.0143	1.7706	1.2884	1.1827	5.0143
Pure Synth.	1.6655	1.1154	0.9918	4.9575	1.6846	1.1512	1.0323	4.9575	1.7047	1.1861	1.0713	4.9575	1.7706	1.2883	1.1825	4.9575
Matching	1.9804	1.2263	1.0740		1.9864	1.2345	1.0819		1.9932	1.2421	1.0887		2.0196	1.2650	1.1075	
Opt. Matching	1.8812	1.3390	1.2200	3.6280	1.8845	1.3451	1.2254	3.4810	1.8894	1.3506	1.2284	3.3950	1.9083	1.3736	1.2457	3.2840
Pen. Synth. (BC)	1.2816	0.5653	0.0127		1.2875	0.5685	0.0259		1.2965	0.5735	0.0400		1.3449	0.6019	0.0909	
Unpen. Synth. (BC)	1.2607	0.5649	0.0152		1.2633	0.5665	0.0176		1.2701	0.5697	0.0207		1.3108	0.5873	0.0288	
Pure Synth. (BC)	1.2607	0.5649	0.0151		1.2634	0.5665	0.0176		1.2702	0.5697	0.0207		1.3108	0.5874	0.0290	
Matching (BC)	1.4708	0.6122	0.0214		1.4741	0.6155	0.0556		1.4817	0.6221	0.0899		1.5267	0.6580	0.1880	
Opt. Matching (BC)	1.3084	0.5818	0.0072		1.3128	0.5851	0.0416		1.3220	0.5910	0.0773		1.3733	0.6336	0.1805	
p = 10, average number of treated outside convex hull: 10.00																
Pen. Synth.	1.8726	1.3592	1.2501	4.6745	1.8981	1.3961	1.2893	4.5992	1.9215	1.4282	1.3231	4.5124	1.9877	1.5135	1.4104	4.2991
Unpen Synth.	1.8494	1.3567	1.2503	5.4268	1.8770	1.3995	1.2971	5.4268	1.9050	1.4410	1.3419	5.4268	1.9926	1.5616	1.4697	5.4268
Pure Synth.	1.8494	1.3567	1.2503	5.4268	1.8770	1.3995	1.2971	5.4268	1.9050	1.4410	1.3419	5.4268	1.9926	1.5616	1.4697	5.4268
Matching	2.1648	1.4732	1.3334		2.1759	1.4867	1.3470		2.1875	1.4993	1.3593		2.2263	1.5360	1.3934	
Opt. Matching	2.0708	1.5651	1.4577	3.6240	2.0802	1.5762	1.4687	3.4790	2.0901	1.5882	1.4806	3.3780	2.1193	1.6236	1.5126	3.2200
Pen. Synth. (BC)	1.3179	0.6141	0.0064		1.3227	0.6167	0.0207		1.3316	0.6211	0.0387		1.3788	0.6485	0.0978	
Unpen. Synth. (BC)	1.2966	0.6077	0.0085		1.2988	0.6077	0.0153		1.3054	0.6098	0.0226		1.3462	0.6260	0.0432	
Pure Synth. (BC)	1.2966	0.6077	0.0085		1.2988	0.6077	0.0153		1.3054	0.6098	0.0226		1.3462	0.6260	0.0432	
Matching (BC)	1.4854	0.6578	0.0057		1.4874	0.6582	0.0425		1.4942	0.6625	0.0795		1.5396	0.6953	0.1854	
Opt. Matching (BC)	1.3534	0.6279	0.0009		1.3573	0.6291	0.0366		1.3659	0.6356	0.0730		1.4120	0.6696	0.1823	

Table 3.3: Monte-Carlo Simulations, $n_1 = 10$, $n_0 = 100$

	$r = 1$				$r = 1.2$				$r = 1.4$				$r = 2$			
	MSE indiv.	MSE mean	|Bias|	Sparsity	MSE indiv.	MSE mean	|Bias|	Sparsity	MSE indiv.	MSE mean	|Bias|	Sparsity	MSE indiv.	MSE mean	|Bias|	Sparsity
$p = 2$, average number of treated outside convex hull: 1.43																
Pen. Synth.	1.2724	0.4375	0.0688	2.5804	1.2711	0.4367	0.0708	2.5783	1.2700	0.4373	0.0715	2.5806	1.2680	0.4373	0.0745	2.5795
Unpen. Synth.	1.1475	0.4517	0.0553	75.6452	1.1504	0.4633	0.1205	75.6452	1.1578	0.4826	0.1815	75.6452	1.1985	0.5668	0.3450	75.6452
Pure Synth.	1.2475	0.4281	0.0647	2.8435	1.2470	0.4285	0.0699	2.8435	1.2466	0.4290	0.0741	2.8435	1.2464	0.4303	0.0838	2.8435
Matching	1.4454	0.4905	0.0816		1.4430	0.4895	0.0786		1.4414	0.4887	0.0758		1.4390	0.4876	0.0687	
Opt. Matching	1.1725	0.4575	0.2407	8.3450	1.1673	0.4538	0.2356	8.4790	1.1618	0.4517	0.2307	8.5130	1.1498	0.4460	0.2271	8.8670
Pen. Synth. (BC)	1.2648	0.4317	0.0236		1.2650	0.4313	0.0235		1.2654	0.4326	0.0226		1.2684	0.4341	0.0205	
Unpen. Synth. (BC)	1.1399	0.4470	0.0169		1.1441	0.4547	0.0806		1.1528	0.4706	0.1400		1.1976	0.5463	0.2990	
Pure Synth. (BC)	1.2405	0.4228	0.0264		1.2411	0.4233	0.0299		1.2420	0.4239	0.0326		1.2455	0.4259	0.0378	
Matching (BC)	1.4264	0.4794	0.0164		1.4267	0.4798	0.0108		1.4277	0.4805	0.0054		1.4328	0.4835	0.0090	
Opt. Matching (BC)	1.1176	0.3791	0.0249		1.1186	0.3782	0.0090		1.1203	0.3801	0.0054		1.1300	0.3863	0.0425	
p = 4, average number of treated outside convex hull: 5.82																
Pen. Synth.	1.2868	0.5276	0.2525	3.6931	1.2821	0.5322	0.2684	3.6955	1.2810	0.5377	0.2818	3.6891	1.2856	0.5562	0.3205	3.6450
Unpen. Synth.	1.2250	0.5073	0.2250	33.7853	1.2269	0.5291	0.2744	33.7853	1.2323	0.5541	0.3219	33.7853	1.2664	0.6410	0.4558	33.7853
Pure Synth.	1.2469	0.5113	0.2326	4.1437	1.2477	0.5232	0.2597	4.1437	1.2499	0.5358	0.2854	4.1437	1.2634	0.5771	0.3572	4.1437
Matching	1.5503	0.6203	0.3304		1.5460	0.6166	0.3248		1.5431	0.6134	0.3194		1.5425	0.6073	0.3067	
Opt. Matching	1.3286	0.6638	0.4936	5.7240	1.3232	0.6602	0.4892	5.6550	1.3171	0.6570	0.4831	5.6230	1.3114	0.6563	0.4770	5.7180
Pen. Synth. (BC)	1.2196	0.4541	0.0202		1.2181	0.4531	0.0152		1.2212	0.4542	0.0125		1.2410	0.4621	0.0072	
Unpen. Synth. (BC)	1.1604	0.4474	0.0238		1.1630	0.4487	0.0156		1.1700	0.4545	0.0533		1.2116	0.4919	0.1596	
Pure Synth. (BC)	1.1835	0.4479	0.0162		1.1849	0.4485	0.0009		1.1885	0.4505	0.0169		1.2085	0.4620	0.0610	
Matching (BC)	1.4308	0.5160	0.0220		1.4332	0.5181	0.0412		1.4378	0.5214	0.0598		1.4618	0.5365	0.1103	
Opt. Matching (BC)	1.1562	0.4231	0.0299		1.1587	0.4282	0.0527		1.1641	0.4345	0.0773		1.1957	0.4608	0.1436	
p = 8, average number of treated outside convex hull: 9.79																
Pen. Synth.	1.5478	0.9490	0.8178	5.0248	1.5631	0.9779	0.8521	4.9698	1.5787	1.0064	0.8834	4.8960	1.6196	1.0743	0.9562	4.6851
Unpen Synth.	1.5221	0.9349	0.8024	6.6565	1.5394	0.9741	0.8489	6.6565	1.5581	1.0126	0.8935	6.6565	1.6231	1.1262	1.0207	6.6565
Pure Synth.	1.5222	0.9352	0.8027	5.6929	1.5395	0.9738	0.8485	5.6929	1.5581	1.0118	0.8924	5.6929	1.6223	1.1236	1.0176	5.6929
Matching	1.8664	1.0824	0.9207		1.8668	1.0849	0.9225		1.8684	1.0873	0.9238		1.8810	1.0964	0.9285	
Opt. Matching	1.7385	1.1782	1.0613	4.3520	1.7385	1.1810	1.0642	4.2450	1.7376	1.1838	1.0656	4.2330	1.7405	1.1895	1.0683	4.1050
Pen. Synth. (BC)	1.1927	0.4670	0.0120		1.1979	0.4686	0.0120		1.2047	0.4731	0.0098		1.2414	0.4932	0.0122	
Unpen. Synth. (BC)	1.1703	0.4655	0.0145		1.1722	0.4671	0.0293		1.1780	0.4705	0.0431		1.2146	0.4888	0.0826	
Pure Synth. (BC)	1.1705	0.4655	0.0147		1.1723	0.4670	0.0289		1.1780	0.4702	0.0420		1.2135	0.4877	0.0795	
Matching (BC)	1.4286	0.5245	0.0046		1.4297	0.5267	0.0282		1.4346	0.5319	0.0608		1.4686	0.5622	0.1524	
Opt. Matching (BC)	1.2186	0.4753	0.0069		1.2228	0.4767	0.0272		1.2276	0.4819	0.0620		1.2664	0.5178	0.1645	
p = 10, average number of treated outside convex hull: 9.97																
Pen. Synth.	1.6937	1.1442	1.0379	5.5262	1.7173	1.1818	1.0793	5.4363	1.7390	1.2146	1.1155	5.3607	1.7991	1.2992	1.2051	5.0745
Unpen Synth.	1.6639	1.1232	1.0205	6.3829	1.6884	1.1678	1.0704	6.3829	1.7138	1.2113	1.1184	6.3829	1.7960	1.3377	1.2551	6.3829
Pure Synth.	1.6640	1.1232	1.0204	6.2551	1.6884	1.1677	1.0703	6.2551	1.7138	1.2111	1.1181	6.2551	1.7959	1.3373	1.2547	6.2551
Matching	2.0378	1.3018	1.1663		2.0431	1.3094	1.1741		2.0490	1.3165	1.1811		2.0731	1.3383	1.2013	
Opt. Matching	1.9132	1.3796	1.2830	3.7810	1.9166	1.3863	1.2898	3.6950	1.9208	1.3930	1.2961	3.6600	1.9408	1.4155	1.3174	3.5530
Pen. Synth. (BC)	1.2011	0.4775	0.0042		1.2059	0.4784	0.0050		1.2129	0.4811	0.0104		1.2540	0.5012	0.0401	
Unpen. Synth. (BC)	1.1796	0.4655	0.0088		1.1812	0.4655	0.0006		1.1865	0.4671	0.0090		1.2215	0.4794	0.0332	
Pure Synth. (BC)	1.1796	0.4656	0.0089		1.1812	0.4656	0.0004		1.1865	0.4672	0.0087		1.2213	0.4795	0.0328	
Matching (BC)	1.4430	0.5533	0.0046		1.4447	0.5550	0.0398		1.4507	0.5600	0.0750		1.4896	0.5917	0.1747	
Opt. Matching (BC)	1.2477	0.4965	0.0026		1.2497	0.5001	0.0345		1.2560	0.5062	0.0706		1.2981	0.5381	0.1748	

Table 3.4: Monte-Carlo Simulations, $n_1 = 100$, $n_0 = 500$

	$r = 1$				$r = 1.2$				$r = 1.4$				$r = 2$			
	MSE indiv.	MSE mean	|Bias|	Sparsity	MSE indiv.	MSE mean	|Bias|	Sparsity	MSE indiv.	MSE mean	|Bias|	Sparsity	MSE indiv.	MSE mean	|Bias|	Sparsity
$p = 2$, average number of treated outside convex hull: 2.14																
Pen. Synth.	1.2272	0.1499	0.0071	2.9346	1.2273	0.1502	0.0089	2.9345	1.3388	0.1590	0.0119	2.9347	1.2273	0.1504	0.0122	2.9341
Unpen. Synth.	1.0806	0.2103	0.0013	210.7918	1.0829	0.2229	0.0724	210.7918	1.0895	0.2517	0.1363	210.7918	1.1240	0.3672	0.2982	210.7918
Matching	1.4153	0.1631	0.0131		1.4149	0.1630	0.0121		1.4147	0.1629	0.0111		1.4146	0.1628	0.0088	
Opt. Matching	1.0558	0.1578	0.0943	14.3700	1.0529	0.1556	0.0908	14.6420	1.0506	0.1534	0.0870	14.8000	1.0452	0.1473	0.0786	15.3160
Pen. Synth. (BC)	1.2267	0.1497	0.0019		1.2269	0.1500	0.0036		1.2387	0.1508	0.0052		1.2274	0.1503	0.0062	
Unpen. Synth. (BC)	1.0800	0.2104	0.0031		1.0825	0.2218	0.0678		1.0892	0.2496	0.1315		1.1242	0.3635	0.2930	
Matching (BC)	1.4119	0.1624	0.0023		1.4122	0.1625	0.0039		1.4126	0.1626	0.0054		1.4142	0.1631	0.0093	
Opt. Matching (BC)	1.0408	0.1245	0.0032		1.0405	0.1255	0.0119		1.0411	0.1268	0.0202		1.0446	0.1334	0.0427	
p = 4, average number of treated outside convex hull: 25.44																
Pen. Synth.	1.1914	0.1833	0.0816	4.4556	1.1918	0.1912	0.0998	4.4450	1.1932	0.1997	0.1160	4.4285	1.1977	0.2243	0.1557	4.3856
Unpen. Synth.	1.1346	0.2062	0.0705	88.2088	1.1374	0.2395	0.1425	88.2088	1.1455	0.2843	0.2099	88.2088	1.1925	0.4366	0.3922	88.2088
Matching	1.4750	0.2454	0.1718		1.4723	0.2424	0.1678		1.4707	0.2397	0.1641		1.4707	0.2342	0.1562	
Opt. Matching	1.1678	0.3724	0.3447	8.8790	1.1612	0.3697	0.3421	9.0500	1.1555	0.3669	0.3392	9.1550	1.1432	0.3643	0.3371	9.7800
Pen. Synth. (BC)	1.1740	0.1622	0.0063		1.1747	0.1624	0.0079		1.1765	0.1638	0.0197		1.1836	0.1720	0.0470	
Unpen. Synth. (BC)	1.1173	0.1927	0.0062		1.1202	0.2027	0.0628		1.1285	0.2315	0.1273		1.1772	0.3605	0.3015	
Matching (BC)	1.4195	0.1733	0.0070		1.4202	0.1742	0.0177		1.4221	0.1759	0.0279		1.4325	0.1836	0.0543	
Opt. Matching (BC)	1.0693	0.1366	0.0047		1.0688	0.1387	0.0230		1.0707	0.1432	0.0405		1.0816	0.1665	0.0897	
p = 8, average number of treated outside convex hull: 85.42																
Pen. Synth.	1.3101	0.5205	0.4874	6.4861	1.3241	0.5655	0.5360	6.4128	1.3386	0.6064	0.5794	6.3324	1.3831	0.6995	0.6766	5.9659
Unpen Synth.	1.2976	0.5086	0.4728	13.2185	1.3127	0.5670	0.5358	13.2185	1.3310	0.6239	0.5961	13.2185	1.4024	0.7874	0.7658	13.2185
Matching	1.7038	0.6995	0.6703		1.7017	0.6978	0.6683		1.7011	0.6960	0.6663		1.7080	0.6941	0.6634	
Opt. Matching	1.4937	0.8550	0.8382	5.0760	1.4891	0.8550	0.8382	5.0500	1.4853	0.8565	0.8396	5.0870	1.4793	0.8615	0.8442	5.1740
Pen. Synth. (BC)	1.1282	0.1746	0.0048		1.1318	0.1775	0.0312		1.1380	0.1846	0.0520		1.1684	0.2055	0.0742	
Unpen. Synth. (BC)	1.1185	0.1800	0.0055		1.1214	0.1871	0.0508		1.1292	0.2038	0.0939		1.1760	0.2844	0.2146	
Matching (BC)	1.4202	0.1896	0.0015		1.4217	0.1918	0.0284		1.4262	0.1979	0.0548		1.4535	0.2311	0.1271	
Opt. Matching (BC)	1.1203	0.1574	0.0012		1.1213	0.1601	0.0290		1.1245	0.1686	0.0593		1.1492	0.2161	0.1444	
p = 10, average number of treated outside convex hull: 95.79																
Pen. Synth.	1.4153	0.7193	0.6950	7.2355	1.4378	0.7732	0.7513	7.1234	1.4607	0.8204	0.8007	6.9857	1.5238	0.9269	0.9099	6.4203
Unpen Synth.	1.4051	0.7064	0.6815	8.8276	1.4277	0.7666	0.7440	8.8276	1.4527	0.8247	0.8040	8.8276	1.5399	0.9905	0.9734	8.8276
Matching	1.8344	0.9291	0.9078		1.8349	0.9308	0.9095		1.8367	0.9322	0.9107		1.8505	0.9385	0.9163	
Opt. Matching	1.6578	1.0750	1.0606	4.3840	1.6566	1.0778	1.0634	4.3250	1.6568	1.0795	1.0649	4.2730	1.6604	1.0915	1.0764	4.2640
Pen. Synth. (BC)	1.1213	0.1768	0.0066		1.1249	0.1764	0.0184		1.1324	0.1786	0.0351		1.1690	0.1937	0.0387	
Unpen. Synth. (BC)	1.1154	0.1805	0.0082		1.1176	0.1828	0.0285		1.1242	0.1921	0.0631		1.1662	0.2465	0.1606	
Matching (BC)	1.4170	0.1897	0.0001		1.4189	0.1922	0.0319		1.4243	0.2001	0.0634		1.4575	0.2449	0.1511	
Opt. Matching (BC)	1.1359	0.1624	0.0001		1.1389	0.1658	0.0344		1.1461	0.1769	0.0687		1.1784	0.2352	0.1661	

returns are defined as the difference between a connected firm's returns and its synthetic control's returns. The measure of the announcement effect is the Cumulative Abnormal Returns (CAR) defined as the sum of abnormal returns since the announcement day.

Our methodology differs in a few ways from the original study, albeit we start with the same base sample. To mitigate complications caused by lack of uniqueness of the synthetic control estimator, Acemoglu et al. (2016) construct synthetic controls on the basis of pretreatment stock returns and restrict the units entering each synthetic control to the 20 untreated units with the highest correlation in returns with the treated unit during the training window. This is a clever ad-hoc solution to the non-uniqueness problem described in Section 1, but it does not easily generalize to contexts where synthetic controls are constructed on the basis of multiple characteristics, and leaves unaddressed the issue of how to decide on the maximum number of units that contribute to the synthetic controls. Instead, we use the full sample of control units and apply the penalized synthetic control estimator proposed in this article, without the pre-selection step in Acemoglu et al. (2016). Moreover, the original study re-weights the CAR of each treated by goodness-of-fit instead of using a simple average across the treated (see equation (7) in their paper). The authors argue that treated firms for which their corresponding synthetic unit better fits its returns over the pre-treatment period should be emphasized because they contain more information. While this assertion makes intuitive sense, especially for cases when a lack of common support prevents a particular treated unit from being well reproduced by a convex combination of control units, the properties of such an estimator are unknown and not covered in the theoretical part of our work.

Table 3.5 reports results. Estimates labeled "corrected inference" discard permuted treated units for which the pre-treatment MSE was three times larger than the mean pre-treatment MSE for the treated units, as in Acemoglu et al. (2016). The results in Table 3.5 are qualitatively similar to the original study, albeit more muted: significance is only obtained at the 5% level in the corrected inference procedure, as compared to significance at the 1% level in the original study. Figure 3.3 displays the Geithner announcement's effect on stock returns versus the Fisher distribution under the no treatment effect assumption. With the selected penalty level of .1, we find that the median number of active controls – defined as having a positive weight in the synthetic unit – for each treated unit is 26.7 (min: 20, max: 40) which is substantially more than in the original analysis where active controls are limited to be 20 or fewer. Another key difference in our inference procedure is that we recompute the cross-validated λ and corresponding synthetic control weights for every member of the treatment group under every permutation, as explained in Section 4. These two observations help explain the difference between our results and the original study.

7.2. The Impact of Election Day Registration on Voter Turnout, Xu (2017)

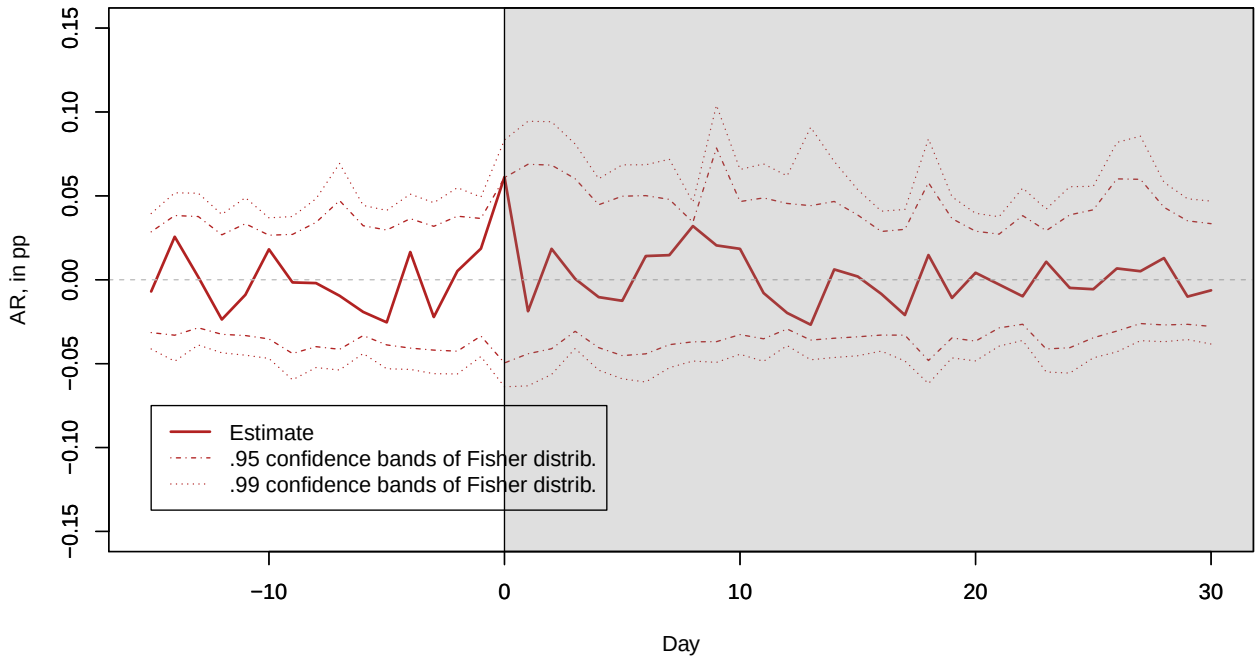
We revisit Xu (2017) which analyzes the impact of Election Day Registration (EDR) on voter turnout in the United States. In most US states, eligible voters must register on a separate day before casting their votes, which entails an extra cost of voting and has been perceived as a cause of low turnout rates. With the objective of raising voter

Table 3.5: Connections to Geithner and Reactions to Treasury Secretary Announcement, Synthetic Control Inference.

	Estimate	Q 0.5%	Q 2.5%	Q 5 %	Q 95%	Q 97.5%	Q 99.5%
				Penalized Synthetic Control			
Day 1, CAR[0,1]	0.061**	- 0.064	- 0.050	- 0.042	0.049	0.061	0.083
	[0.008; 0.116]						
Day 10, CAR[0,10]	0.138*	- 0.128	- 0.093	- 0.075	0.126	0.150	0.202
	[0.016; 0.261]						
				Corrected Inference			
Day 1, CAR[0,1]	0.061**	- 0.065	- 0.049	- 0.042	0.045	0.058	0.087
Day 10, CAR[0,10]	0.138*	- 0.123	- 0.091	- 0.073	0.116	0.142	0.202
				Bias-corrected Estimator			
Day 1, CAR[0,1]	0.058	- 0.108	- 0.080	- 0.067	0.064	0.077	0.105
Day 10, CAR[0,10]	0.125	- 0.229	- 0.171	- 0.142	0.156	0.186	0.247
Cross-val. (MSE) λ	0.08						
Mean Sparsity	26.7						
				Non-Penalized Synthetic Control, $\lambda = 0$			
Day 1, CAR[0,1]	0.060**	- 0.070	- 0.054	- 0.046	0.047	0.060	0.082
Day 10, CAR[0,10]	0.114*	- 0.155	- 0.124	- 0.108	0.094	0.119	0.171
				Corrected Inference			
Day 1, CAR[0,1]	0.060**	- 0.068	- 0.053	- 0.045	0.044	0.057	0.087
Day 10, CAR[0,10]	0.114*	- 0.165	- 0.126	- 0.111	0.084	0.114	0.171
				Bias-corrected Estimator			
Day 1, CAR[0,1]	0.058	- 0.110	- 0.082	- 0.068	0.063	0.076	0.104
Day 10, CAR[0,10]	0.119	- 0.238	- 0.178	- 0.150	0.149	0.180	0.243
Mean Sparsity	40.8						
Sample size (n)	525						
Nb. in treatment group (n_1)	12						

Note: This table displays Cumulative Abnormal Returns (CAR) on day 1 and 10 corresponding to panels B and C, columns 2 and 3, of Table 5 in Acemoglu et al. (2016). Results are obtained on their base sample which excludes the 10% firms whose returns are most correlated with Citigroup. We define being treated as at least one meeting between the firm and Geithner in 2007-08. The estimate column corresponds to the difference between the treated returns and synthetic control returns accumulated for the said number of days since announcement. The number between brackets are Fisher confidence intervals at 95% levels, based on 5,000 permutations, computed by inverting the tests. The quantiles displayed in the other columns are computed as quantiles of the Fisher distribution under the no-effect assumption. 20,000 random permutations have been used. Corrected inference discards permuted treated units for which the pre-treatment MSE was three times larger than the mean pre-treatment MSE for the treated units, as in Acemoglu et al. (2016). Bias-corrected inference relies on a linear specification for the regression function. Asterisks denote significance levels (** = 5%, * = 10%).

Figure 3.3: Abnormal Returns after Geithner Announcement, non-corrected inference



Note: The confidence bands are computed as quantiles of the Fisher distribution under the sharp null hypothesis of no treatment effect. They do not define a confidence interval for the treatment effect. When the solid red line goes out of these bands, it means the effect is significant. 20,000 permutations are used. The shadowed grey area is the post-announcement period.

turnout, EDR laws were first implemented in Maine, Minnesota and Wisconsin in 1976; Idaho, New Hampshire and Wyoming followed suit in 1996; finally, Montana, Iowa and Connecticut adopted the legislation as well in 2012. We refer the reader to the original article for further details on the policy. The dataset, available in the `R` package `gsynth` (Xu and Liu, 2018), comprises turnout rates measured during 24 elections (from 1920 to 2012), for 47 US states among which 9 are treated (*i.e.* adopted the EDR) and 38 are non-treated. Since the adoption of the treatment was staggered, we consider first the nine treated states together with treatment starting in 1976 and then break down the analysis for each wave of adoption.

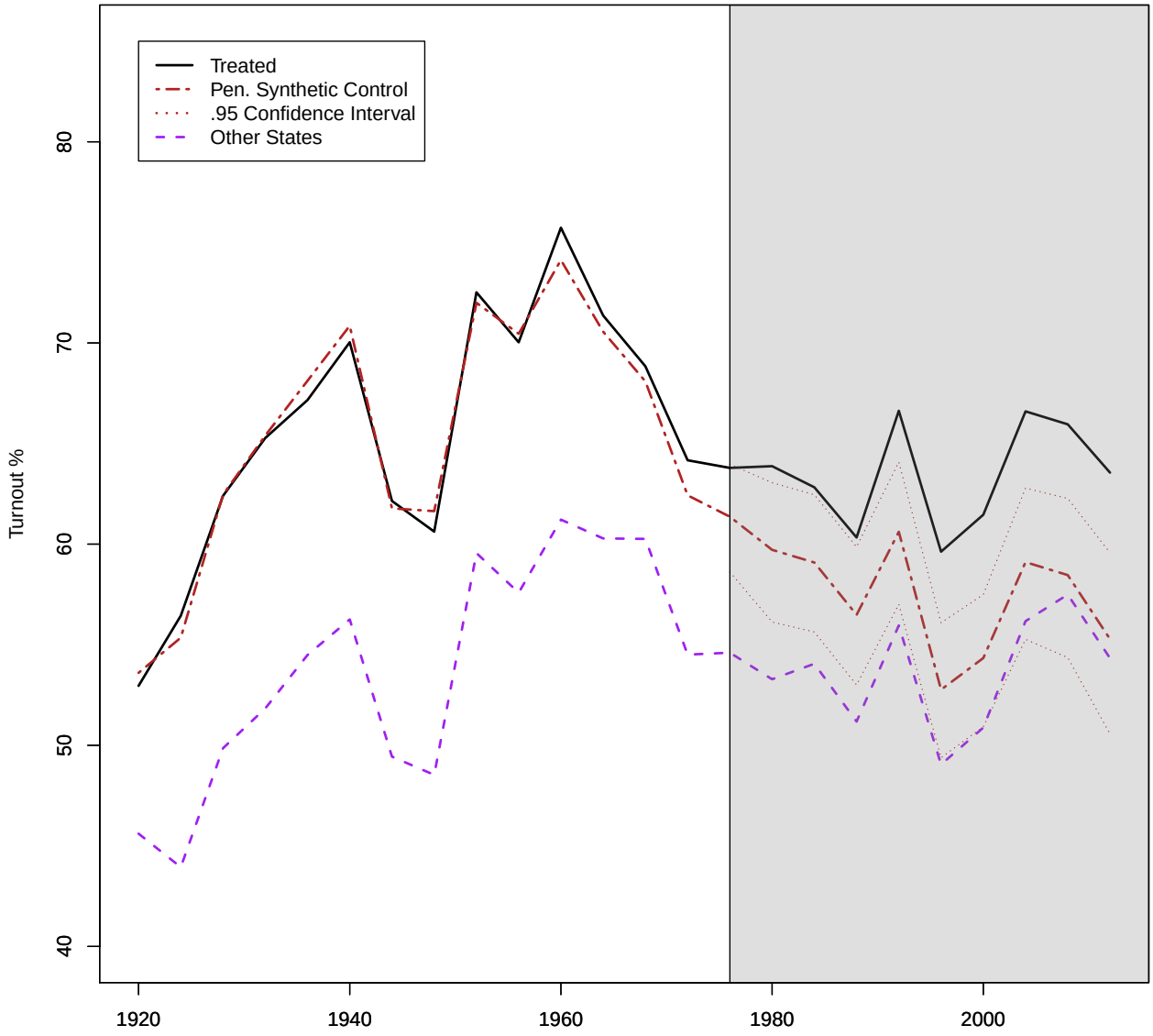
To choose λ we employ the pre-intervention holdout procedure of Section 5.2. The training sample uses turnout rates over the elections 1920-1948. The validation sample uses elections 1952-1972. The procedure ends up selecting a very low value of .004 for λ, so the penalized synthetic control estimate is very close to the non-penalized one. Figure 3.4 displays actual and counterfactual turnout rates computed using the penalized synthetic control estimator. The impact is positive and significant at 5% for every post-treatment election. The p-values obtained from randomized inference ($B = 10,000$) using the MSPE ratio (3.10) and the sum of ranks for the individual effects aggregated over the post-treatment period are 5×10^{-3} and 2×10^{-4} respectively (see Section 4 for more details). Due to the dimensionality of the problem ($p = 14$), all the treated are outside of the convex hull defined by the untreated, so the pure synthetic control estimator is equal to the non-penalized synthetic control estimator. Furthermore, all the treated are connected to all the untreated in the augmented Delaunay tesselation.

We use a similar strategy when breaking down the results by wave of adoption, except that for the second and third waves, more pre-treatment periods are available to select the optimal λ and construct the counterfactual. Elections 1976-1992 and 1996-2008 are further available for the second and third waves, respectively. We select a $\lambda \approx 0$ for the first wave (6 to 8 non-zero untreated units per synthetic unit) and $\lambda \approx .5$ for the second (3 to 4 non-zero untreated units per synthetic unit) and third (2 untreated units per synthetic unit). Figure 3.5 breaks down the results for each wave. Our analysis confirms the original study of Xu (2017), by finding that results are mainly driven by the first-adopters while the adoption of EDR is statistically insignificant at the 5% level for the states who adopted it later.

8. Conclusion

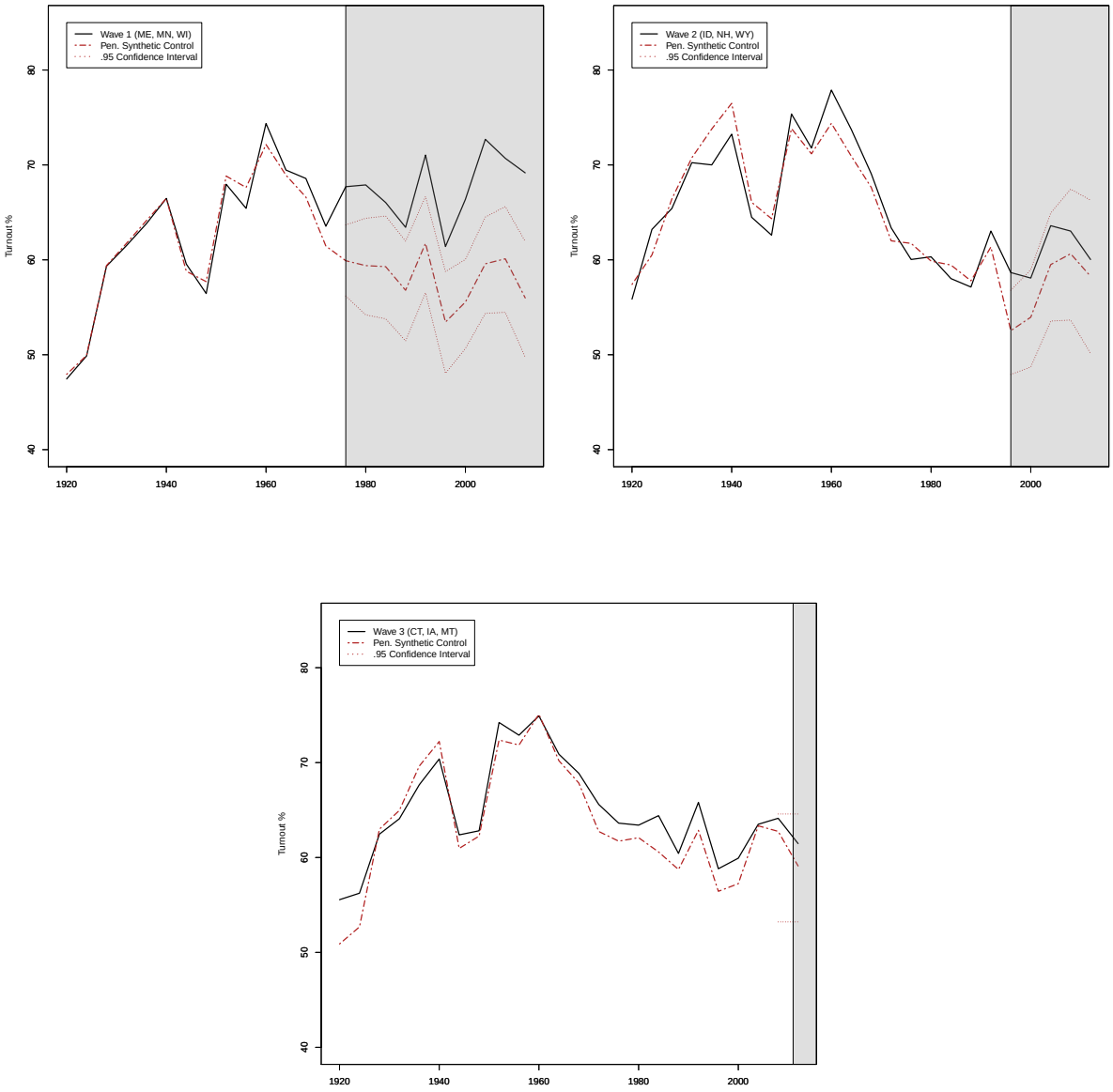
In this chapter, we proposed a penalized synthetic control estimator that trades-off pairwise matching discrepancies with respect to the characteristics of each unit in the synthetic control against matching discrepancies with respect to the characteristics of the synthetic control unit as a whole. We studied the properties of this estimator and proposed data driven choices of the penalization parameter. We showed that the penalized synthetic control estimator is unique and sparse, which makes it particularly convenient for empirical applications with many treated units, where the focus is on average treatment effects. Motivated by the case with many treated and untreated sample units, we derived large sample properties of the penalized synthetic control estimator and proposed

Figure 3.4: Voter Turnout in the US and EDR Laws



Note: The .95 confidence intervals are computed by inverting Fisher Tests. 10,000 permutations are used. The dashed purple line is the average turnout per election for the 38 nontreated States.

Figure 3.5: Voter Turnout in the US and EDR Laws, by Wave of Adoption



Note: The top left panel is the first wave of adoption (ME, MN, WI); the top right is the second (ID, NH, WY); the bottom is the third (CT, IA, MT). The .95 confidence intervals are computed by inverting Fisher Tests. 10,000 permutations are used.

a bias-correction as in Abadie and Imbens (2011). The penalized synthetic control estimators perform well in simulations. Finally, we illustrated their practical applicability in two empirical examples based on Acemoglu et al. (2016) and Xu (2017).

9. Appendix: Proofs

Notation

For any real matrix X, let $\mathcal{CH}(X)$ and $\mathcal{DT}(X)$ be the convex hull and the Delaunay tesselation of the columns of X, respectively. We recall that $\mathcal{DT}(X)$ is a partition of $\mathcal{CH}(X)$. μ^{Leb} denotes the Lebesgue measure.

Proof of Lemma 3.1

Notice that if the first result in Lemma 3.1 does not hold, then $W_i^*(\lambda)$ cannot be a solution to the problem in equation (3.6). We start by proving the upper bound in the second inequality. Since $W_i^*(\lambda)$ minimizes (3.6), it follows that

$$(X_i - X_0 W_i^*(\lambda))'(X_i - X_0 W_i^*(\lambda)) + \lambda \Delta_i' W_i^*(\lambda) \leq (X_i - X_{NN_i})'(X_i - X_{NN_i}) + \lambda \Delta_i^{NN}$$

Therefore,

$$\lambda \Delta_i' W_i^*(\lambda) \leq (1+\lambda)\Delta_i^{NN},$$

and the result follows from $\lambda > 0$. The lower bound follows from the definition of Δ_i^{NN}. □

Proof of Theorem 3.1

Without loss of generality, consider the case with only one treated, $n_1 = 1$. Program (3.8) is

$$\begin{aligned} \min_{W} \quad & f_\lambda(W) = (X_1 - X_0 W)'(X_1 - X_0 W) + \lambda W' \Delta_1, \\ \text{s.t.} \quad & W \in \mathcal{W}, \end{aligned} \tag{C.1}$$

where $\mathcal{W} = \{W \in [0,1]^{n_0} \,|\, W' 1_{n_0} = 1\}$. It is easy to check that the feasible set, $\mathcal{W}$, is convex and compact. Because f_λ is continuous and $\mathcal{W}$ is compact, it follows that the function attains a minimum on $\mathcal{W}$. Moreover, $X_0' X_0$ is positive semi-definite, so f_λ is convex.

Suppose that more than one solution exist. In particular, assume that W_1 and W_2 are solutions, with $f_\lambda(W_1) = f_\lambda(W_2) = f_\lambda^*$. Then, for any $a \in (0,1)$ we have that $aW_1 + (1-a)W_2 \in \mathcal{W}$. Because f_λ is convex, we obtain

$$f_\lambda(aW_1 + (1-a)W_2) \leq a f_\lambda(W_1) + (1-a) f_\lambda(W_2) = f_\lambda^*.$$

This implies that the problem has either a unique solution or infinitely many. In addition, if there are multiple solutions they all produce the same fitted values $X_0 W$. To prove this suppose there are two solutions W_1 and W_2 such that $X_0 W_1 \neq X_0 W_2$. Then, because $\|x - c\|^2$ is strictly convex in c, for $a \in (0,1)$ we obtain

$$\begin{aligned} f_\lambda(aW_1 + (1-a)W_2) &= \|X_1 - X_0(aW_1 + (1-a)W_2)\|^2 + \lambda(aW_1 + (1-a)W_2)'\Delta_1 \\ &< a\|X_1 - X_0 W_1\|^2 + (1-a)\|X_1 - X_0 W_2\|^2 + \lambda(aW_1 + (1-a)W_2)'\Delta_1 \\ &= a f_\lambda^* + (1-a) f_\lambda^* \end{aligned}$$

$$= f^*_\lambda,$$

which contradicts that W_1 and W_2 are solutions. As a result, if W_1 and W_2 are solutions, then $X_0W_1 = X_0W_2$. Moreover, $\lambda > 0$ implies $W_1'\Delta_1 = W_2'\Delta_1$. Let $A = [X_0' \ 1_{n_0} \ \Delta_1]$. It follows that, if W_1 and W_2 are solutions, then $A'(W_1 - W_2) = 0_{p+2}$ (where 0_{p+2} is a $(p+2) \times 1$ vector of zeros).

Karush-Kunh-Tucker conditions imply:

$$X_j'(X_1 - X_0W) - \frac{\lambda}{2}\Delta_{1,j} = \pi - \gamma_j$$
$$W_j \geq 0, \ W'1_{n_0} = 1, \ \gamma_j \geq 0, \ \gamma_j W_j = 0.$$

Stacking the first n_0 conditions above and pre-multiplying by W', we obtain

$$W'X_0'(X_1 - X_0W) - \frac{\lambda}{2}W'\Delta_1 = \pi.$$

From this equation, it follows that the value of π is unique across solutions, because $X_0'W$ and $W'\Delta_1$ are unique across solutions. Given that π is unique, the equations

$$X_j'(X_1 - X_0W) - \frac{\lambda}{2}\Delta_{1,j} = \pi - \gamma_j.$$

imply that the γ_j's are unique across solutions. Let $\widetilde{X}_0$ be the submatrix of X_0 formed by the columns associated with zero γ_j's, and define $\widetilde{W}$, $\widetilde{\Delta}_1$, and $1_{\widetilde{n}_0}$ analogously, where $\widetilde{n}_0$ is the number of columns of $\widetilde{X}_0$. Then,

$$\widetilde{X}_0'(X_1 - \widetilde{X}_0\widetilde{W}) = \frac{\lambda}{2}\widetilde{\Delta}_1 + \pi 1_{\widetilde{n}_0}. \tag{C.2}$$

Notice that if $\lambda > 0$, then $\|X_1 - X_0W\| = 0$ implies that $\widetilde{\Delta}_1$ is a constant vector. We therefore obtain that if $\lambda > 0$ and $\widetilde{\Delta}_1$ is not constant, then it must be the case that $\|X_1 - X_0W\| > 0$.

Let $\widetilde{A} = [\widetilde{X}_0' \ 1_{\widetilde{n}_0} \ \widetilde{\Delta}_1]$. Consider the case $\widetilde{n}_0 \geq p+2$. In this case $\widetilde{A}$ has full column rank, which implies that equation (C.2) cannot hold if $\lambda > 0$. As a result, when $\lambda > 0$, the solution to (C.1) has $p+1$ non-zero components at most.

Consider now the case $\widetilde{n}_0 \leq p+1$. For this case $\widetilde{A}$ has full row rank. Moreover, if $\widetilde{W}_1$ and $\widetilde{W}_2$ are solutions, it must be the case that $\widetilde{A}'(\widetilde{W}_1 - \widetilde{W}_2) = 0_{p+2}$. However, because $\widetilde{A}$ has full row rank the system $\widetilde{A}'z = 0_{p+2}$ admits only the trivial solution, $z = 0_{\widetilde{n}_0}$, which implies that the solution to (C.1) is unique. □

Lemma 3.4 (Optimality of Delaunay for the Compound Discrepancy, Rajan, 1994) *Let $Z \in \mathcal{CH}(X_0)$. Consider a solution $\widetilde{W} = (\widetilde{W}_{n_1+1}, \ldots, \widetilde{W}_n)'$ of the problem*

$$\min_{W \in [0,1]^{n_0}} \sum_{j=n_1+1}^{n} W_j \|X_j - Z\|^2, \tag{C.3}$$

$$\text{s.t.} \quad X_0 W = Z, \quad \sum_{j=n_1+1}^{n} W_j = 1. \tag{C.4}$$

Then, non-zero values of $\widetilde{W}_j$ occur only among the vertices of the face of the Delaunay complex containing Z.

We restate the proof of Lemma 10 in Rajan (1994) for clarity and note that it does no rely on general quadratic position of the set of points.

Proof of Lemma 3.4

For a point $X \in \mathbb{R}^p$, consider the transformation $\phi : X \to (X, \|X\|^2)$. The images under ϕ of points in $\mathbb{R}^p$ belong to the paraboloid of revolution $\mathcal{P}$ with vertical axis and equation $x_{p+1} = \sum_{i=1}^{p} x_i^2$. By Theorem 17.3.1 in Boissonnat and Yvinec (1998), the faces of the Delaunay complex of the n_0 points $X_{n_1+1}, \ldots, X_n$ in $\mathbb{R}^p$ are obtained by projecting onto $\mathbb{R}^p$ the faces of the lower envelope of the convex hull of the n_0 points $\phi(X_{n_1+1}), \ldots, \phi(X_n)$ obtained by lifting the X_j's onto the paraboloid $\mathcal{P}$.

Now consider points $\left(\sum_{j=n_1+1}^{n} W_j X_j, \sum_{j=n_1+1}^{n} W_j \|X_j\|^2\right)$ subject to the constraints in (C.4). These points are equal to $\left(Z, \sum_{j=n_1+1}^{n} W_j \|X_j - Z\|^2 + \|Z\|^2\right)$ and belongs to the convex hull of $\phi(X_{n_1+1}), \ldots, \phi(X_n)$. Hence, a solution of (C.3) for a fixed Z is given by such a point with the lowest $(p+1)$-th coordinate. It is a point on the lower envelope of the convex hull of $\phi(X_{n_1+1}), \ldots, \phi(X_n)$, so Z belongs to a p-face of the Delaunay complex. As a consequence, the only non-zero entries of $\widetilde{W}$ occur only among the vertices of the face of the Delaunay complex of the columns of X_0 containing Z. □

Proof of Theorem 3.2

It is enough to prove that the result holds for one treated unit, so we consider the case $n_1 = 1$ and drop the treated units subscripts from the notation. We proceed by contradiction. Suppose that the synthetic control weights are given by the vector $W^*(\lambda) = (W_2^*(\lambda), \ldots, W_n^*(\lambda))'$, and that $W_j^*(\lambda) > 0$ for j which is not a vertex of the face of the Delaunay complex $\mathcal{DT}(X_0)$ containing $X_0 W^*(\lambda)$. Because $X_0 W^*(\lambda) \in \mathcal{CH}(X_0)$, it follows from Lemma 3.4 that we can always choose an n_0-vector of weights $\widetilde{W} \in [0,1]^{n_0}$, such that *(i)* $X_0 \widetilde{W} = X_0 W^*(\lambda)$, *(ii)* $\sum_{j=2}^{n} \widetilde{W}_j = 1$, *(iii)* $\widetilde{W}_j = 0$ for any j that is not a vertex of the face of the Delaunay complex containing $X_0 W^*(\lambda)$, and *(iv)* $\widetilde{W}$ induces a lower compound discrepancy than $W^*(\lambda)$ relative to $X_0 \widetilde{W} = X_0 W^*(\lambda)$,

$$\sum_{j=2}^{n} \widetilde{W}_j \|X_j - X_0 \widetilde{W}(\lambda)\|^2 < \sum_{j=2}^{n} W_j^*(\lambda) \|X_j - X_0 W^*(\lambda)\|^2. \tag{C.5}$$

For any $W \in [0,1]^{n_0}$ it can be easily seen that

$$\sum_{j=2}^{n} W_j \|X_j - X_1\|^2 = \sum_{j=2}^{n} W_j \|X_j - X_0 W\|^2 + \|X_1 - X_0 W\|^2. \tag{C.6}$$

Combining equations (C.5) and (C.6) with the fact that $\|X_1 - X_0\widetilde{W}\|^2 = \|X_1 - X_0 W^*(\lambda)\|^2$, we obtain

$$\sum_{j=2}^{n} \widetilde{W}_j \|X_j - X_1\|^2 < \sum_{j=2}^{n} W_j^*(\lambda) \|X_j - X_1\|^2.$$

As a result

$$\|X_1 - X_0\widetilde{W}\|^2 + \lambda \sum_{j=2}^{n} \widetilde{W}_j \|X_j - X_1\|^2 < \|X_1 - X_0 W^*(\lambda)\|^2 + \lambda \sum_{j=2}^{n} W_j^*(\lambda) \|X_j - X_1\|^2,$$

which contradicts the premise that $W^*(\lambda)$ is a solution to (3.6). □

Proof of Theorem 3.3

Since the columns of $[X_1\ X_0]$ are in general quadratic position, the augmented Delaunay triangulation, $\mathcal{DT}([X_1 : X_0])$, exists and is unique. Without loss of generality, consider the case of a single treated unit, and normalize X_1 to be at the origin. Let X_0^* be the submatrix of X_0 formed by the columns of X_0 that are connected to X_1 in the augmented triangulation, $\mathcal{DT}([X_1 : X_0])$, and $\mathcal{UT}(X_1, X_0)$ be the union of the Delaunay simplices that have X_1 as a vertex in $\mathcal{DT}([X_1 : X_0])$. Consider a point $z \in \mathcal{CH}([X_1 : X_0]) \backslash \mathcal{UT}(X_1, X_0)$. We will first show that z cannot be equal to $X_0 W^*(\lambda)$. Because z does not belong to $\mathcal{UT}(X_1, X_0)$ and because the set $\mathcal{CH}([X_1 : X_0])$ is convex, it is always possible to find a point $v \in \mathcal{CH}([X_1 : X_0]) \backslash \mathcal{UT}(X_1, X_0)$ on the line segment that connects z and X_1 such that $\|X_1 - v\| < \|X_1 - z\|$ (or, equivalently, $\|v\| < \|z\|$). For any point in $x \in \mathcal{CH}([X_1 : X_0])$ consider the set of non-negative weights, $w_1(x), \ldots, w_n(x)$, such that: *(i)* $\sum_{i=1}^{n} w_i(x) = 1$, *(ii)* $\sum_{i=1}^{n} w_i(x) X_i = x$, and *(iii)* if X_i is not a vertex of the Delaunay simplex containing x, then $w_i(x) = 0$. If $x \in \mathcal{CH}([X_1 : X_0]) \backslash \mathcal{UT}(X_1, X_0)$, then the Delaunay simplex containing x in $\mathcal{DT}([X_1 : X_0])$ is the same as the the Delaunay simplex contain x in $\mathcal{DT}(X_0)$ (Devillers and Teillaud, 2003; Boissonnat et al., 2009). Therefore, by Theorem 3.2, if $x \in \mathcal{CH}([X_1 : X_0]) \backslash \mathcal{UT}(X_1, X_0)$ and $X_0 W^* = x$, then $W^*(\lambda) = (w_2(x), \ldots, w_n(x))'$. Now, let $f(x) = \sum_{i=1}^{n} w_i(x) \|X_i\|^2 = \sum_{i=1}^{n} w_i(x) \|X_1 - X_i\|^2$. This function is convex because it is the lower boundary of the convex hull of $\{(X_1, \|X_1\|^2), \ldots, (X_n, \|X_n\|^2)\}$ (Rajan, 1994), and is minimized at $x = X_1$. As we move from z to v we travel in the direction of the minimun of $f(x)$. Because $f(x)$ is a convex function, it follows that $f(v) < f(z)$. Because $\|v\| < \|z\|$ and $f(v) < f(z)$, it follows that $X_0 W^*(\lambda) \neq z$, regardless of the value of λ. This implies that $X_0 W^*(\lambda)$ must belong to $\mathcal{UT}(X_1, X_0)$ and the result follows from Theorem 2. □

Lemma 3.5 (Sum of Weights) *For $j = n_1 + 1, ..., n$, denote $S_j(\lambda) = \sum_{i=1}^{n_1} W_{i,j}^*(\lambda)$, the sum of weights given to a particular control unit across all the synthetic units. Under Assumption 3.1, for any $\lambda \geq 0$:* (i) $\sum_{j=n_1+1}^{n} S_j(\lambda) = n_1$, (ii) $E[S_j(\lambda)] = n_1/n_0$ *for every $j = n_1 + 1, ..., n$, and* (iii) $\rho(S_j(\lambda), S_k(\lambda)) = -1/(n_0 - 1)$ *for any $j \neq k$, where* $\rho(S_j(\lambda), S_k(\lambda)) = \text{cov}(S_j(\lambda), S_k(\lambda))/\text{var}(S_j(\lambda))$.

Proof of Lemma 3.5

The first assertion holds because there are n_1 synthetic units, so summing the share of all synthetic units generated from every donor must yield n_1. The second assertion is a

consequence of the previous one, the linearity of the expectation operator and exchangeability. For the third assertion, notice that the first statement of the lemma implies $\text{var}(\sum_{j=n_1+1}^{n} S_j(\lambda)) = 0$ which in combination with exchangeability leads to:

$$n_0 \text{var}(S_j(\lambda)) + n_0(n_0 - 1)\text{cov}(S_j(\lambda), S_k(\lambda)) = 0. \tag{C.7}$$

A consequence of equation (C.7) is that $\rho(S_j(\lambda), S_k(\lambda)) = \text{cov}(S_j(\lambda), S_k(\lambda))/\text{var}(S_j(\lambda)) = -1/(n_0 - 1)$. □

Proof of Lemma 3.2

The bias term of the estimator is a simple average of the individual bias terms: $B_n(\lambda) = n_1^{-1} \sum_{i=1}^{n_1} b_i(\lambda)$, where the individual bias term for treated unit i is defined as $b_i(\lambda) := \mu_0(X_i) - \sum_{j=n_1+1}^{n} W_{i,j}^*(\lambda)\mu_0(X_j)$. Notice that because the synthetic weights sum to one:

$$b_i(\lambda) = \sum_{j=n_1+1}^{n} W_{i,j}^*(\lambda)[\mu_0(X_i) - \mu_0(X_j)].$$

From Assumption 3.4, $\mu_0(.)$ is Lipschitz-continuous with constant C_{μ_0} and using Jensen's inequality, keeping in mind that $\sum_{j=n_1+1}^{n} W_{i,j}^*(\lambda) = 1$

$$\begin{aligned}
|b_i(\lambda)| &\leq C_{\mu_0} \sum_{j=n_1+1}^{n} W_{i,j}^*(\lambda)\|X_i - X_j\| \\
&\leq C_{\mu_0} \sqrt{\sum_{j=n_1+1}^{n} W_{i,j}^*(\lambda)\|X_i - X_j\|^2} \\
&\leq C_{\mu_0} \sqrt{\frac{1+\lambda}{\lambda}\Delta_i^{NN}},
\end{aligned}$$

where the last inequality uses Lemma 3.1. The aggregated bias term is therefore bounded:

$$|B_n(\lambda)| \leq C_{\mu_0} \sqrt{\frac{1+\lambda}{\lambda}} \frac{1}{n_1} \sum_{i=1}^{n_1} \sqrt{\Delta_i^{NN}}.$$

Using the previous inequality and Jensen's inequality again, notice

$$\begin{aligned}
E\left[n^{2/p} B_n^2(\lambda)\right] &\leq n^{2/p} C_{\mu_0}^2 \frac{1+\lambda}{\lambda} E\left[\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\sqrt{\Delta_i^{NN}}\right)^2\right] \\
&\leq C_{\mu_0}^2 \frac{1+\lambda}{\lambda} E\left[n^{2/p}\Delta_1^{NN}\right] \\
&< \infty,
\end{aligned}$$

where the last inequality follows from Lemma 2 in Abadie and Imbens (2006). Now, Chebyshev's inequality implies the result of Lemma 3.2. □

Proof of Theorem 3.4

Lemma 3.2 implies $B_n(\lambda) \xrightarrow{p} 0$. Next, we will show that $M_n(\lambda) \xrightarrow{p} 0$. Since *(i)* the support of $X|D=1$ is contained in the support of $X|D=0$, which is bounded, and *(ii)* μ_0 is Lipschitz, we obtain $E[\mu_0(X_i)|D_i=1] < \infty$. Also $|\mu_1(X_i)| = |E[Y_i|X_i, D_i=1]| \leq E\big[|Y_i|\big|X_i, D_i=1\big]$. As a result, $E\big[|\mu_1(X_i)|\big|D_i=1\big] \leq E[|Y_i||D_i=1] < \infty$. By the weak law of large numbers, we obtain

$$\frac{1}{n_1}\sum_{i=1}^{n_1}(\mu_1(X_i) - \mu_0(X_i) - \tau) \xrightarrow{p} 0,$$

and

$$\frac{1}{n_1}\sum_{i=1}^{n_1}(Y_i - \mu_1(X_i)) \xrightarrow{p} 0.$$

By Chebyshev's inequality, for any $\varepsilon > 0$:

$$\begin{aligned}
&\Pr\left(\left|\frac{1}{n_1}\sum_{i=n_1+1}^{n} S_i(\lambda)\big(Y_i - \mu_0(X_i)\big)\right| > \varepsilon\right) \\
&\qquad \leq \frac{1}{\varepsilon^2}\frac{1}{n_1^2}E\left[\left(\sum_{i=n_1+1}^{n} S_i(\lambda)\big(Y_i - \mu_0(X_i)\big)\right)^2\right] \\
&\qquad = \frac{1}{\varepsilon^2}\frac{1}{n_1^2}E\left[\sum_{i=n_1+1}^{n} S_i(\lambda)^2\big(Y_i - \mu_0(X_i)\big)^2\right] \\
&\qquad + \frac{1}{\varepsilon^2}\frac{1}{n_1^2}E\left[2\sum_{i=n_1+1}^{n}\sum_{j>i}^{n} S_i(\lambda)S_j(\lambda)\big(Y_i - \mu_0(X_i)\big)\big(Y_j - \mu_0(X_j)\big)\right] \\
&\qquad = \frac{1}{\varepsilon^2}\frac{n_0}{n_1^2}E\Big[S_{n_1+1}^2(\lambda)\big(Y_{n_1+1} - \mu_0(X_{n_1+1})\big)^2\Big] \\
&\qquad \leq \frac{1}{\varepsilon^2}\bar{\sigma}_0^2\frac{n_0}{n_1^2}E[S_{n_1+1}^2(\lambda)],
\end{aligned}$$

where the second equality follows from $E[S_i(\lambda)S_j(\lambda)(Y_i - \mu_0(X_i))(Y_j - \mu_0(X_j))|D_1, \ldots, D_n, X_1, \ldots, X_n] = 0$ for $n_1 + 1 \leq i < j \leq n$. Now, the result follows from Assumption 3.5. □

In the proof of Theorem 3.5 we will use the notation $(Y_{n,i}, D_{n,i}, X_{n,i})$, for the outcome, treatment, and covariates of observation i at sample size n, respectively. Similarly, $S_{n,i}(\lambda)$ is the sum of the weights assigned to observation i, for i such that $D_{n,i} = 0$. In this notation, the order of the observations is invariant in n (i.e., there is no reordering). For ease of reference we reproduce next a version of a Martingale CLT with random norming (Hall and Heyde, 1980, Theorem 3.3, p. 64). We will use this result in the proof of Theorem 3.5.

Theorem 3.6 (Martingale Central Limit Theorem) *Let* $\{T_{n,k}, \mathcal{F}_{n,k}, 1 \leq k \leq k_n, n \geq 1\}$

be a zero-mean, square-integrable martingale array with differences $\xi_{n,k}$ *and squared variations* $U^2_{n,k_n} = \sum_{k=1}^{k_n} \xi^2_{n,k}$ *and let* η^2 *be an a.s. finite random variable. Suppose that as* $n \to \infty$

$$\sum_{k=1}^{k_n} E[\xi^2_{n,k} \mathbf{1}\{|\xi_{n,k}| > \varepsilon\} | \mathcal{F}_{n,k-1}] \xrightarrow{p} 0, \text{ for all } \varepsilon > 0, \tag{C.8}$$

$$V^2_{n,k_n} = \sum_{k=1}^{k_n} E[\xi^2_{n,k} | \mathcal{F}_{n,k-1}] \xrightarrow{p} \eta^2, \tag{C.9}$$

with $\Pr(\eta^2 > 0) = 1$*, and the* σ*-fields are nested* $\mathcal{F}_{n,k} \subset \mathcal{F}_{n+1,k}$ *for* $1 \leq k \leq k_n, n \geq 1$*. Then:*

$$T_{n,k_n}/U_{n,k_n} = \frac{\sum_{k=1}^{k_n} \xi_{n,k}}{\sqrt{\sum_{k=1}^{k_n} \xi^2_{n,k}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Proof of Theorem 3.5

Consider the martingale array $\sqrt{n_1} M_n(\lambda) = \sum_{k=1}^{2n} \xi_{n,k}$, where

$$\xi_{n,k} = \begin{cases} \frac{1}{\sqrt{n_1}} D_{n,k} \left(\mu_1(X_{n,k}) - \mu_0(X_{n,k}) - \tau\right) & \text{if } 1 \leq k \leq n, \\ \frac{1}{\sqrt{n_1}} \left(D_{n,k-n} - (1 - D_{n,k-n}) S_{n,k-n}(\lambda)\right) \left(Y_{n,k-n} - \mu_{D_{n,k-n}}(X_{n,k-n})\right) & \text{if } n+1 \leq k \leq 2n. \end{cases}$$

Let $\mathcal{X}_n = \{X_{n,1}, ..., X_{n,n}\} = \{X_1, ..., X_n\}$, $\mathcal{D}_n = \{D_{n,1}, ..., D_{n,n}\} = \{D_1, ..., D_n\}$ and consider $\mathcal{F}_{n,k} = \sigma(\mathcal{D}_n, X_{n,1}, ..., X_{n,k})$ for $k \leq n$ and $\mathcal{F}_{n,k} = \sigma(\mathcal{D}_n, \mathcal{X}_n, Y_{n,1}, ..., Y_{n,k-n})$ for $k > n$. We will verify the conditions of Theorem 3.6. The Lyapunov's condition

$$\sum_{k=1}^{2n} E[|\xi_{n,k}|^{2+\delta} | \mathcal{F}_{n,k-1}] \xrightarrow{p} 0,$$

for some $\delta > 0$ is sufficient for (C.8). For the first n terms of the martingale array,

$$\sum_{k=1}^{n} E[|\xi_{n,i}|^{2+\delta} | \mathcal{F}_{n,k-1}] = \frac{1}{n_1^{\delta/2}} E[|\mu_1(X) - \mu_0(X) - \tau|^{2+\delta} | D = 1] \to 0.$$

Now consider the last n terms. For $\delta > 0$, let $C_\delta < \infty$ be a bound on $E[|Y - \mu_D(X)|^{2+\delta} | X = x, D = d]$. Then,

$$\begin{aligned} &\sum_{k=n+1}^{2n} E[|\xi_{n,k}|^{2+\delta} | \mathcal{F}_{n,k-1}] \\ &= \frac{1}{n_1^{1+\delta/2}} \sum_{k=n+1}^{2n} |(D_{n,k-n} - (1 - D_{n,k-n}) S_{n,k-n}(\lambda)|^{2+\delta} E[|Y_{n,k-n} - \mu_{D_{n,k-n}}(X_{n,k-n})|^{2+\delta} | \mathcal{F}_{n,k-1}] \\ &\leq \frac{C_\delta}{n_1^{\delta/2}} \frac{1}{n_1} \sum_{k=n+1}^{2n} |D_{n,k-n} - (1 - D_{n,k-n}) S_{n,k-n}(\lambda)|^{2+\delta} \xrightarrow{p} 0. \end{aligned}$$

Now, given that

$$E\left[\frac{1}{n_1}\sum_{k=n+1}^{2n}|D_{n,k-n}-(1-D_{n,k-n})S_{n,k-n}(\lambda)|^{2+\delta}\right]=1+\frac{n_0}{n_1}E[S^{2+\delta}_{n,n_1+1}(\lambda)]<\infty,$$

by Markov's inequality we obtain $\sum_{k=n+1}^{2n}E[|\xi_{n,k}|^{2+\delta}|\mathcal{F}_{n,k-1}]\xrightarrow{p}0$.
Turn to (C.9). Because of exchangeability, the conditional variance is

$$V^2_{n,2n}=E\left[(\mu_1(X)-\mu_0(X)-\tau)^2\,\middle|D=1\right]+\frac{1}{n_1}\sum_{i=1}^{n_1}\sigma_1^2(X_i)+\frac{1}{n_1}\sum_{j=n_1+1}^{n}S_{n,j}(\lambda)^2\sigma_0^2(X_j).$$

Part (i) of Assumption 3.6 ensures that the middle term converges to $E[\sigma_1^2(X)|D=1]$ by the weak law of large numbers. Focusing on the last term, Chebychev's inequality together with part (iv) of Assumption 3.6 imply that for some $\varepsilon>0$:

$$\begin{aligned}
&Pr\left(\left|\frac{1}{n_0}\sum_{j=n_1+1}^{n}S_{n,j}(\lambda)^2\sigma_0^2(X_j)-E\left[S_{n,n_1+1}(\lambda)^2\sigma_0^2(X_{n_1+1})\right]\right|>\varepsilon\right)\\
&\quad\leq\frac{1}{n_0}\frac{var[S_{n,n_1+1}(\lambda)^4\sigma_0^4(X_{n_1+1})]}{\varepsilon^2}+\frac{n_0(n_0-1)}{n_0^2}\frac{cov\left[S_{n,n_1+1}(\lambda)^2\sigma_0^2(X_{n_1+1}),S_{n,n_1+2}(\lambda)^2\sigma_0^2(X_{n_1+2})\right]}{\varepsilon^2}\\
&\quad\leq\frac{1}{n_0}\frac{E\left[S_{n,n_1+1}(\lambda)^4\sigma_0^4(X_{n_1+1})\right]}{\varepsilon^2}+\frac{cov\left[S_{n,n_1+1}(\lambda)^2\sigma_0^2(X_{n_1+1}),S_{n,n_1+2}(\lambda)^2\sigma_0^2(X_{n_1+2})\right]}{\varepsilon^2}\\
&\quad\leq\frac{1}{n_0}\frac{\bar\sigma_0^4E\left[S_{n,n_1+1}(\lambda)^4\right]}{\varepsilon^2}+\frac{cov\left[S_{n,n_1+1}(\lambda)^2\sigma_0^2(X_{n_1+1}),S_{n,n_1+2}(\lambda)^2\sigma_0^2(X_{n_1+2})\right]}{\varepsilon^2}\\
&\quad\to 0.
\end{aligned}$$

As a consequence, as $n_1,n_0\to\infty$, we obtain that the conditional variance converges to a positive limit:

$$\begin{aligned}
V^2_{n,2n}\xrightarrow{p}&E\left[(\mu_1(X)-\mu_0(X)-\tau)^2\,\middle|D=1\right]+E\left[\sigma_1^2(X)\middle|D=1\right]\\
&+\lim_{n\to\infty}\frac{n_0}{n_1}E\left[S_{n,n_1+1}(\lambda)^2\sigma_0^2(X_{n_1+1})\right].
\end{aligned}$$

Applying Theorem 3.6:

$$\widehat\sigma^{-1}(\lambda)\sqrt{n_1}M_n(\lambda)\xrightarrow{d}\mathcal{N}(0,1),\text{ as }n_1,n_0\to\infty,$$

where

$$\widehat\sigma^2(\lambda):=\frac{1}{n_1}\sum_{i=1}^{n}D_i\left(\mu_1(X_i)-\mu_0(X_i)-\tau\right)^2+\frac{1}{n_1}\sum_{i=1}^{n}D_i\sigma_1^2(X_i)+(1-D_i)S_i(\lambda)^2\sigma_0^2(X_i).$$

□

Lemma 3.6 (Conditional Probability of a Link) *Under Assumptions 3.1, 3.2 and 3.7, for any $i = 1, \ldots, n_1$, $j = n_1 + 1, \ldots, n$, for $\lambda > 0$ and any integer $r > 0$*

$$E[\Pr\left(W^*_{i,j}(\lambda) > 0|X_0\right)^r |D = 0] = \mathcal{O}\left([\log(n_0)/n_0]^r\right).$$

Proof of Lemma 3.6

It is enough to prove that the result holds for one non-treated unit and for one treated unit, so we set $i = 1$. Theorem 3.3 establishes the necessary condition that control unit j has to be linked to the treated unit through an edge in $\mathcal{DT}([X_1 : X_0])$ for $W^*_{1,j}(\lambda) > 0$ to occur. We can rely on the incremental algorithm to construct Delaunay tessellations to make this condition more practical in probabilistic terms (Berg et al., 2008, Chapter 9). Start from $\mathcal{DT}(X_0)$ and insert X_1 to create the augmented Delaunay tessellation $\mathcal{DT}([X_1 : X_0])$. Theorem 9.7 in Berg et al. (2008) implies that any simplex in $\mathcal{DT}(X_0)$ whose circumscribed hypersphere does not contain X_1 remains a Delaunay simplex and will belong to $\mathcal{DT}([X_1 : X_0])$, hence the insertion of X_1 only destroys (in the Delaunay sense) simplices whose circumscribed hypersphere contains X_1. Two cases have to be distinguished:

(i) if X_1 falls in a circumscribed hypersphere, then X_1 has to be connected to all of the vertices of the corresponding simplex,

(ii) if X_1 does not fall in any circumscribed hypersphere, $\mathcal{DT}(X_0) \subset \mathcal{DT}([X_1 : X_0])$ and p new edges are created. In this case, X_1 necessarily falls outside of the convex hull $\mathcal{CH}(X_0)$ of the columns of X_0.

Define F_j, the *flower* of X_j, as the union of the circumscribed hyperspheres of Delaunay simplices of $\mathcal{DT}(X_0)$ which have X_j as a vertex. From the discussion above, if $X_1 \in F_j$ then X_1 is linked to X_j in the new tessellation $\mathcal{DT}([X_1 : X_0])$. Now, let us decompose the event "X_1 linked to X_j" by whether or not X_1 falls into $\mathcal{CH}(X_0)$:

$$\begin{aligned}
\Pr\left(W^*_{1,j}(\lambda) > 0|X_0\right) \leq & \Pr\left(X_1 \text{ linked to } X_j|X_0\right) \\
= & \Pr\left(X_1 \text{ linked to } X_j \cap X_1 \in \mathcal{CH}(X_0)|X_0\right) \\
& + \Pr\left(X_1 \text{ linked to } X_j \cap X_1 \notin \mathcal{CH}(X_0)|X_0\right) \\
\leq & \Pr\left(X_1 \in F_j \cap X_1 \in \mathcal{CH}(X_0)|X_0\right) + \Pr\left(X_1 \notin \mathcal{CH}(X_0)|X_0\right) \\
\leq & \Pr\left(X_1 \in F_j|X_0\right) + \Pr\left(X_1 \notin \mathcal{CH}(X_0)|X_0\right).
\end{aligned}$$

Using Minkowski's inequality we obtain:

$$\begin{aligned}
E[\Pr\left(W^*_{i,j}(\lambda) > 0|X_0\right)^r |D = 0] \leq & E[\Pr\left(X_1 \in F_j|X_0\right)^r |D = 0] \\
& + E[\Pr\left(X_1 \notin \mathcal{CH}(X_0)|X_0\right)^r |D = 0). \qquad \text{(C.10)}
\end{aligned}$$

Notice that $\Pr\left(X_1 \notin \mathcal{CH}(X_0)|X_0\right) = \Pr\left(X_1 \in \mathcal{X} \setminus \mathcal{CH}(X_0)|X_0\right) \leq \bar{f}_1 \mu^{Leb}(\mathcal{X} \setminus \mathcal{CH}(X_0))$. In order to show that $\mu^{Leb}(\mathcal{X} \setminus \mathcal{CH}(X_0)) \xrightarrow{p} 0$, we use Theorem 2 in Brunel (2017).

Part (i) of Assumption 3.7 ensures that the so-called *margin condition* is satisfied with $\alpha = 0$ and that the density of X_0 is bounded from above. As a consequence, the *missing volume*, $\mu^{Leb}(\mathcal{X} \setminus \mathcal{CH}(X_0)) \xrightarrow{p} 0$ as $n_0 \to \infty$. Convergence also happens in L_r, $r \geq 1$, since $\Pr(X_1 \notin \mathcal{CH}(X_0)|X_0) \leq 1$ (see also Corollary 2 in Brunel, 2017).

Focus on $\Pr(X_1 \in F_j|X_0)$. F_j is contained in a hypersphere $\mathcal{C}$ centered at X_j of radius that is twice the length of the radius of the largest circumscribed hypersphere of a Delaunay simplex that has X_j as a vertex. Define an *empty hypersphere* with respect to X_0 as a sphere that does not contain any column of X_0. By definition, all hyperspheres that circumscribe a Delaunay simplex in $\mathcal{DT}(X_0)$ are empty hyperspheres with respect to X_0. As a consequence, the volume of $\mathcal{C}$ is less than the volume of a hypersphere of radius twice that of the largest empty hypersphere. Denote V_{n_0} the volume of the largest empty hypersphere with respect to X_0. Mathematically, we have

$$\Pr(X_1 \in F_j|X_0) \leq \overline{f_1}\mu^{Leb}(F_j) \leq \overline{f_1}2^p V_{n_0},$$

which yields:

$$E\left[\Pr(X_1 \in F_j|X_0)^r|D=0\right] \leq \overline{f_1}^r 2^{pr} E\left[V_{n_0}^r|D=0\right].$$

Define the random variable $Z_{n_0} := n_0 V_{n_0} - \log(n_0) - (p-1)\log(\log(n_0)) - \log(\gamma)$ where

$$\gamma := \frac{1}{p!}\left[\sqrt{\pi}\frac{\Gamma(p/2+1)}{\Gamma((p+1)/2)}\right]^{p-1}.$$

We have:

$$E\left[V_{n_0}^r|D=0\right] = \frac{1}{n_0^r}E\left[(Z_{n_0} + \log(n_0) + (p-1)\log(\log(n_0)) + \log(\gamma))^r|D=0\right].$$

Under parts (i) and (iii) of Assumption 3.7, Theorem 2 in Aaron et al. (2017) states that $Z_{n_0} \xrightarrow{d} U$, as $n_0 \to \infty$, where $P[U \leq u] = \exp(-\exp(-u))$. The random variable U has finite moments of order r that do not depend on n_0, which proves that $E[\Pr(X_1 \in F_j|X_0)^r|D=0] = \mathcal{O}([\log(n_0)/n_0]^r)$. In light of equation (C.10), we have proven the result. □

Proof of Lemma 3.3

It is enough to prove that the result holds for one non-treated unit. Define $B_{n_1+1}(\lambda) := \sum_{i=1}^{n_1} \mathbf{1}\{W^*_{i,n_1+1}(\lambda) > 0\}$. $S_{n_1+1}(\lambda)^2 \leq B_{n_1+1}(\lambda)$ a.s. Treated units $X_1, \ldots, X_{n_1}$ are independent by Assumption 3.1 so conditionally on $X_{n_1+1}, \ldots, X_n$, weights $W^*_{1,n_1+1}(\lambda), \ldots, W^*_{n_1,n_1+1}(\lambda)$ are independent. They are also identically distributed, so $B_{n_1+1}(\lambda)|X_0 \sim \mathcal{B}\left(n_1, \Pr\left(W^*_{1,n_1+1}(\lambda) > 0|X_0\right)\right)$ For $m \geq 1$, the m-th moment of $B_{n_1+1}(\lambda)$ conditional on X_0 is

$$E[B_{n_1+1}(\lambda)^m|X_0] = \sum_{k=0}^{m} S(m,k)\frac{n_1!\Pr\left(W^*_{1,n_1+1}(\lambda) > 0|X_0\right)^k}{(n_1-k)!}$$

$$\leq \sum_{k=0}^{m} S(m,k) n_1^k \Pr\left(W^*_{1,n_1+1}(\lambda) > 0 | X_0\right)^k,$$

where $S(m,k)$ are Stirling numbers of the second kind. Then, because for $m \geq 1$, $S(m,0) = 0$ and $S(m,1) = 1$:

$$\begin{aligned}
\frac{n_0}{n_1^2} E\left[S_{n_1+1}(\lambda)^m | D = 0\right] &\leq \frac{n_0}{n_1^2} E\left[B_{n_1+1}(\lambda)^m | D = 0\right] \\
&= \frac{n_0}{n_1^2} E\left[E\left[B_{n_1+1}(\lambda)^m | X_0\right] | D = 0\right] \\
&\leq \frac{n_0}{n_1^2} \sum_{k=1}^{m} S(m,k) n_1^k E\left[\Pr\left(W^*_{1,n_1+1}(\lambda) > 0 | X_0\right)^k | D = 0\right] \\
&\leq \frac{n_0}{n_1^2} \left(\frac{n_1(p+1)}{n_0} + \sum_{k=2}^{m} S(m,k) n_1^k E\left[\Pr\left(W^*_{1,n_1+1}(\lambda) > 0 | X_0\right)^k | D = 0\right] \right) \\
&\leq \frac{p+1}{n_1} + C \sum_{k=2}^{m} S(m,k) \left(\frac{n_1}{n_0}\right)^{k-2} \frac{\log(n_0)^k}{n_0},
\end{aligned}$$

for some positive constant C. The last inequality uses Lemma 3.6. In light of the assumption that n_1/n_0 is bounded, the right-hand side of the last inequality goes to zero. □

Chapter 4

Using Generic Machine Learning to Analyze Treatment Heterogeneity: An Application to Provision of Job Counseling

Joint work with Bruno Crépon, Esther Duflo and Elia Pérennès.

Summary

We re-analyze a large-scale randomized experiment conducted in France in 2007-2008 that was created to evaluate the impact of an intensive job-search counseling program on employment outcomes. This experiment was specifically designed to compare public and private provisions of job counselling. Originally, Behaghel et al. (2014) found the public program to be twice as effective as the private program, a finding that they partially blame on the payment structure to which the private providers were subject. We strongly suspect that private providers had incentives to accept candidates without any regards for their expected treatment effect. In particular, it is likely that they cream-skimmed candidates based on their expected baseline probability of finding a job and that they also indulged in parking candidates (*i.e.* maxing-out the number of trainees but providing minimal training). Using the generic machine learning methodology developed in Chernozhukov et al. (2018b), we are able to find evidence of heterogeneous treatment effects and construct ML proxies for the individualized treatment effect. We then study the impact of both the individualized treatment effect and the individualized baseline probability of finding a job on the likelihood of enrollment in both programs.

1. Introduction

Who benefits from a treatment? What is the optimal treatment assignment? How are applicants to Randomized Controlled Trials (RCT) selected when the treatment is not mandatory? These are classical questions in empirical Economics that call for prediction of the individual treatment effect. Fortunately, the developments of supervised Machine Learning (ML) in the last decades have provided high-quality tools for prediction. However, applying them to predict individual treatment effects is not completely straightforward because the outcome under both treatment status is never observed for the same individual, the individual treatment effect is never observed which means this is not a simple supervised learning task, and since these tools have been designed for prediction and not so much is known about their theoretical properties, they need to be adapted to answer causal inference questions (Athey, 2015) . Recently, a burgeoning econometric literature has tried to bridge that gap and integrate ML to the applied researcher's toolbox (*e.g.* Belloni et al., 2014b; Chernozhukov et al., 2018a; Athey and Wager, 2018; Athey and Imbens, 2019; Chernozhukov et al., 2018b to cite only a few). Besides the technical aspects, the introduction of ML into empirical Economics has opened a new exciting avenue of research that studies the potential improvement of human decisions by machine predictions, *e.g.* Kleinberg et al. (2017).

With these questions in mind, we revisit a large-scale randomized experiment conducted in France in 2007-2008 that was created to evaluate the impact of an intensive job-search counseling program on employment outcomes. This experiment was specifically designed to compare a public and a private provisions of job counselling. Originally, Behaghel et al. (2014) found the public program to be twice as effective as the private program. They attribute this discrepancy to insufficient mastering of the counselling technology on the private arm of the treatment and potential incentive problems in the design of the contracts, particularly in a context where there exist heterogeneous propensities to exit unemployment in the population. In particular, all private providers were subject to a two-part payment structure: 30 percent of the maximal sum when the job seeker enrolled intro the program, and 70 percent conditional on placement (35 percent if the job was found within six months and the other 35 percent if the worker was still employed after six months). They suspect that this payment structure entailed two types of side effects: if the fixed part of the payment is large, private providers are likely to maximize enrollment intro the program and offer very little job counselling to keep the costs down (*parking*); if the conditional payment is relatively large, they are likely to enroll the candidates with the best labor market prospects and again, provide them with little job counselling (*cream-skimming*). On the other hand, the public arm of the program provided by the French Public Employment Service (PES), was not subject to financial incentives.

The goal of this article is to study the selection process, by the caseworkers, of the program applicants. In particular, we suspect that private providers indulged in parking and/or cream-skimming. For that, we use ML algorithms to predict the baseline probability of finding a job and the treatment effect at the individual level. We adapt the Generic Machine Learning framework developed by Chernozhukov et al. (2018b) to randomized controlled trials with imperfect compliance where the target parameter of interest is the

Local Average Treatment Effect (LATE). Put succinctly, the Generic Machine Learning framework allows to study heterogeneity in the effect of the treatment as long as the econometrician can produce two quantities: a ML proxy of the treatment effect, *i.e.* a prediction of the individual treatment effect made using a machine learning algorithm, and an unbiased signal of the individual treatment effect, *i.e.* a random variable such that its expectation conditional on observable characteristics is equal to the individual treatment effect. They develop a test that allows to rule out the absence of heterogeneity and estimate some features of the distribution of the Conditional Average Treatment Effect (CATE). Contrary to standard RCTs with perfect compliance, a unbiased signal for the LATE does not exist and complicates the straightforward adaption of this convenient framework.

Section 2 draws an overview of the use of ML tools in empirical Economics. Section 3 describes the data and the experimental design. Section 4 deals with the empirical strategy the we follow, while highlighting both the economic mechanisms we have in mind when studying the selection process and the adaptation of the generic machine learning framework to this context. Section 5 displays the results. Section 6 concludes.

2. Machine Learning in Empirical Economics

Taking its roots in the post-selection inference problem (Leamer, 1983; Leeb and Pötscher, 2005), the integration of machine learning tools to the empirical economist's toolkit started with the Lasso. Likely spurred by the good theoretical understanding of this method, it has been studied in a number of settings and models relevant for practitioners, be it instrumental variables (Belloni et al., 2012), panel data (Belloni et al., 2016), demand estimation (Chernozhukov et al., 2017a), discriminations (Bach et al., 2018), among others.

Many contributions in this literature were quick to highlight the potential benefits of these modern statistical tools for policy evaluation and causal inference, while acknowledging the difficulties posed by adapting them to achieve goals that are standard in empirical Economics and which are often broader than prediction (*e.g.* Athey, 2015; Athey and Imbens, 2016). At first, the perceived value-added of machine learning methods relied mostly in variable selection and high-quality estimation of high-dimensional nuisance parameters under an unconfoundedness assumption, *e.g.* Belloni et al. (2014b,a, 2017); Farrell (2015). Notice that while these methods had been applied in several empirical studies, they were not suited for inference until a few years ago. Recent contributions went beyond the Lasso to integrate other non-standard statistical tools of sufficiently high-quality to the empiricist's inference toolbox, such as random trees and forests, boosting, support vector machines, kernel methods, or neural networks (*e.g.* Chernozhukov et al., 2018a). Two ingredients played a key role in this breakthrough. First, the use of orthogonal scores ensures that the resulting treatment effect estimator is first-order insensitive to deviations from the true value of nuisance parameters (Chernozhukov et al., 2015; Chernozhukov et al., 2015; Chernozhukov et al., 2018). In other words, orthogonal scores mitigate the impact of replacing unknown nuisance parameters by machine learning estimators that are often not $\sqrt{n}$-consistent. Most of the time, such a strategy requires the estimation of

more nuisance parameters, yielding the name of *double selection* (Belloni et al., 2014a) or *double machine learning* (Chernozhukov et al., 2017; Chernozhukov et al., 2018a). Second, sample-splitting appeared as a way to limit the proclivity of these tools to over-fit the data. Indeed, several papers (*e.g.* Athey and Wager, 2018) advocate for splitting the data between an *auxiliary* sample where nuisance parameters are estimated and a *main* sample where the parameter of interest is estimated using out-of-sample predictions based on the predictors constructed on the auxiliary sample. Chernozhukov et al. (2017, 2018a) suggest that the role of the two samples be then switched and the estimators combined in order to prevent loss of efficiency, yielding the name *cross-fitting*.

Analysis of RCTs was, by design, less likely to benefit from the use of more complex econometric tools. Indeed, independence between the treatment variable and confounding factors immunizes RCT from a selection bias, granting them the status of 'gold standard' of scientific evidence among regularly studied designs (*e.g.* Abadie and Cattaneo, 2018). By leveraging the prediction performance of machine learning tools, inference regarding treatment effect in RCT can, however, benefit from an increase in precision (lower standard error). Furthermore, they offer a way of searching for treatment effect heterogeneity in the absence of a pre-analysis plan because they provide a flexible way to estimate the Conditional Average Treatment Effect (CATE), that is, the expected treatment effect conditional on some individual characteristics. For example, by their own natures, random trees and random forests partition the data to predict the outcome of interest (they find thresholds and interact variables in a data-driven fashion, Mullainathan and Spiess, 2017) and can be adapted for causal inference purposes (see the *causal trees* of Athey and Wager, 2018). Contrary to pre-analysis plans, they don't require to specify the dimensions along which the economist will search for heterogeneity beforehand while still not allowing for p-hacking if done correctly (Chernozhukov et al., 2018b). Indeed, pre-analysis plans can be costly because they are inflexible and end up wasting a lot of data points (Olken, 2015). In other words, machine learning 'lets the data speak' and allows to discover dimensions along which the treatment effect differs even if they previously were not suspected to matter.

So far, applications of these modern statistical methods to RCT have been scarce. Davis and Heller (2017b) and Davis and Heller (2017a) use causal forests to study the heterogeneity in effectiveness of a youth summer job program in reducing the probability of committing crimes and increasing the likelihood of attending school or being employed. Davis and Heller (2017a) successfully identify a sub-group for which the program increases employment while the effect of the program for employment is not statistically significant on average. This sub-group appears to differ from the youth usually targeted by these programs, thereby questioning the state of knowledge in the field.

3. Data and Experimental Design

We analyze a large-scale randomized experiment conducted in France in 2007-2008 and designed to evaluate the impact of an intensive job-search counseling on employment outcomes. An initial evaluation can be found in Behaghel et al. (2014).

3.1. Design of the Experiment

We study a program dedicated to help job seekers. This program involves the provision of personalized job-search assistance : the job-seeker is assigned a dedicated personal advisor with a much lower caseload than in the standard track. This translates into at least one weekly contact (by e-mail or telephone) and one monthly face-to-face meeting between the job-seeker and the caseworker. The average caseload ratio in this program is around 40 job-seekers per caseworker. Compared to the usual track, where a contact is supposed to take place every month and where PES agents assist on average 120 job-seekers, this is a significant increase in the human resources dedicated to assisting the job-seeker. Participation in the program is voluntary. Job-seekers are enrolled by signing a charter and their 6-month trajectory within the program is organized around an individual action plan, the objectives of which are periodically reviewed.

The main steps of the experimental protocol consist in identifying a set of jobseekers eligible for intensive job-search counseling and then to draw between three modalities of accompaniment : standard track provided by the PES, intensive job-search counseling provided by a private provider, intensive job-search counseling provided by the PES. There are three categories of eligible job-seekers : the newly unemployed (< 3 months) entitled to at least one year of benefits, the newly unemployed (< 3 months) entitled to benefits for less than one year and the long-term unemployed (> 3 months). The identification of eligible job-seekers is done by the local agency advisor, during an interview with the job-seeker. The counselor identifies whether the job-seeker is part of the public for whom the intensive support program was planned (the long-term unemployed) and follows a set of statistical criteria (risk of long-term unemployment) and more qualitative criteria (distance from employment, a sufficiently defined professional project). Probabilities of assignment to each group varied locally and across time so as to maximize the statistical power of the evaluation while complying with the quantitative objectives of the program. The random assignment took place over 15 months, from January 2007 to March 2008, in 393 local public employment offices in 16 of the 22 French administrative regions.

Both the public and the private intensive job-search programs have previously been evaluated by Behaghel et al. (2014). They found that job-search assistance increases exit rates to employment from 15 to 35% and that the impact of the public program was about twice as large as that of the private program, at least during the first 6 months after random assignment. Finally, they found that the effect of the public program was relatively homogeneous with respect to gender, education, and age.

3.2. Data

Our sample consists of all newly unemployed individuals (for less than 3 months at the time of randomization) with sufficient benefit entitlement and who were subject to the random assignment for participation in the program. The total sample size is $57,661$ individuals, from which $40,373$ individuals were assigned to the private program, $7,345$ individuals were assigned to the public program and the others in the control group. Our analysis is based on an administrative data file on job seekers provided by the Public

Employment Service. These administrative records provide sociodemographic information on job-seekers (age, gender, education level, family situation, reason for registration, start and end dates of unemployment spells, level of UB, hours of work while on claim, etc.). Some of the pre-treatment variables are continuous, some are categorical and some are binary.

As noted above, program participation is voluntary. Job-seekers assigned to more intensive programs are more likely to enter that program but they do not systematically comply with the assignment: on average, program participation is around 46% among those encouraged to participate to the private program and around 45% among those encouraged to participate to the public program (see Table 4.1).

Table 4.1: Program participation

		Entry				
		Standard	Public		Standard	Private
Assignation	Standard	96.98%	3.02%	Standard	95.11%	4.89%
	Public	55.22%	44.78%	Private	54.28%	45.72%

Table 4.10 in appendix provides some descriptive statistics among experimental groups.

4. Empirical Strategy

4.1. An Economic Model of Treatment Allocation

Behaghel et al. (2014) found that private program was half as effective as the public program. Three reasons may be invoked: (i) the private providers did not master the counselling technology as well as the PES, (ii) the populations enrolled in the public and private arms of the treatment were different, and (iii) the private providers may have been incentivized to enroll job-seekers and then provide little-to-no effort. We study these last two possibilities. All private providers were subject to a two-part payment structure: 30 percent of the maximal sum when the job-seeker enrolled intro the program, and 70 percent conditional on placement (35 percent if the job was found within six months and the other 35 percent if the worker was still employed after six months). This payment structure may entail two types of side effects: if the fixed part of the payment is large, private providers are likely to maximize enrollment into the program and offer very little job counselling to keep the costs down (*parking*); if the conditional payment is relatively large, they are likely to target the candidates with the best labor market prospect and again, provide little job counselling (*cream-skimming*). On the other hand, the public arm of the program provided by the French Public Employment Service (PES), was not subject to financial incentives.

We suppose that the treatment choice is up to the training providers rather than to the individual candidate once the randomization has taken place. Each job-seeker is defined by the triple (Y_0, Y_1, X) where Y_0 and Y_1 are the probability of finding of job under the standard track and under the treatment, respectively, and X is a set of observable

characteristics. Only X is observable by the PES and the private providers so we assume that they base their decisions on the couple $(\mu_0(X), \tau(X))$ where $\mu_0(X) = E[Y_0|X]$ and $\tau(X) = E[Y_1 - Y_0|X]$. For simplicity, we assume that $(\mu_0(X), \tau(X))$ is the same for both arms of the treatment and study the estimates of the treatment effect we will obtain in both cases. In the following, we start from the base model in Section 5 of Behaghel et al. (2012).

The PES will provide job-seekers with job counselling for those such that the benefits of the treatment (the unemployment benefits U times the increase in the probability of finding a job) will be larger than its cost (c_{pub}): $\tau(X)U \geq c_{pub}$. As a consequence, the optimal treatment rule is to treat all job-seekers such that their individual treatment effect is larger than the relative cost of treatment, c_{pub}/U and we obtain $E[\tau(X)|\text{Public}] = E[\tau(X)|\tau(X) \geq c_{pub}/U]$. In this context, and because the program was specifically designed to have case-workers in charge of a moderate number of job-seekers, there is no incentive for the PES to park candidates. On the other hand, if $\tau(X)$ and $\mu_0(X)$ are positively correlated, it may be optimal to cream-skim candidates, or, at least, following that optimal treatment rule may look like cream-skimming in the data.

The private provider is subject to a two-part payment structure: for a maximal payment of P, $(1-\delta)P$ is paid on enrollment of a job-seeker and δP is paid conditional on placement. In the context of the experiment, $\delta = .7$ and $P \in [3,000; 4,000]$ (euros). Within this incentive scheme, the private case-worker has two decisions to take: (i) which job-seekers should be accepted into the program? and (ii) which job-seekers should be actually treated, *i.e.* who should the case-worker spend effort on? Let us start with the second question and assume that the pure cost of enrollment is $\underline{c}$, while the cost (including enrollment) of treatment for a job-seeker is $c_{pri} > \underline{c}$. If the private case-worker decides to not exert any effort on the job-seeker, its expected profit is $\Pi_0 = (1-\delta)P + \mu_0(X)\delta P - \underline{c}$. If the private case-worker decides to exert an effort on the job-seeker, its expected profit is $\Pi_1 = (1-\delta)P + (\mu_0(X) + \tau(X))\delta P - c_{pri}$. As a consequence, a job-seeker is enrolled if $\min(\Pi_0, \Pi_1) \geq 0$, that is to say, either if it has baseline labor-market prospects $\mu_0(X)$ larger than the threshold $[\underline{c}/P - (1-\delta)]/\delta$, or if it is a "high enough"-responder to treatment, *i.e.* $\mu_0(X) + \tau(X) \geq [c_{pri}/P - (1-\delta)]/\delta$. Notice that cream-skimming is directly a feature of the model, and if the threshold $[\underline{c}/P - (1-\delta)]/\delta$ is low enough, it may not even qualify as cream-skimming in the sense that any job-seeker is welcomed in the program (but may not be treated). If that is the case, more job-seekers than is optimal can end up being treated. Moreover, a job-seeker gets treated if and only if $\Pi_1 \geq \Pi_0$, that is to say if the job-seeker is a high-responder $\tau(X) \geq (c_{pri} - \underline{c})/\delta P$. The larger the cost of effort, the more likely the private provider is to indulge in parking. If the cost of effort is too large that is, larger than c_{pub}/U, job-seekers that would benefit from the treatment under the PES would go untreated when assigned to the private program. Notice also that if expectations on the treatment effect, are harder to form *i.e.* $\tau(X)$ is a weak signal of $Y_1 - Y_0$, while $\mu_0(X)$ is a stronger signal of Y_0, the rule $\tau(X) \geq (c_{pri} - \underline{c})/\delta P$ may oftentimes be violated, giving a low estimate of the treatment effect when considering the private program.

4.2. Methodological Aspects

We are interested in performing inference over the Local Average Treatment Effect (LATE) conditional on covariates. However, adapting the framework of Chernozhukov et al. (2018b) to randomized experiments with imperfect compliance is not straightforward.

Let Y_0 and Y_1 denote the potential outcomes under no-treatment and treatment, respectively. Let D_0 and D_1 denote the potential treatments under non-assignment and assignment, respectively. Let X be a vector of observed covariates and Z a binary variable coding for the treatment assignment. We adopt the ubiquitous assumptions of the LATE framework (see for example Assumption 2.1 in Abadie, 2003). For each individual in the sample, we observe the vector (Y, D, Z, X) where $D = D_0 + Z(D_1 - D_0)$ and $Y = Y_0 + D_Z(Y_1 - Y_0)$. Consider the following structural model:

$$\begin{aligned} Y &= \mu(X) + \tau(X)D + \varepsilon, \\ D &= \alpha(X) + \beta(X)Z + U, \end{aligned}$$

where we assume that $E[\varepsilon|X, Z] = E[U|X, Z] = 0$, but in general $E[\varepsilon|D, X] \neq 0$ because treatment choice is endogeneous. Notice that the second equation yields $\alpha(X) = E[D_0|X]$ and $\beta(X) = E[D_1 - D_0|X] = P[D_1 > D_0|X]$, the probability of being a complier conditional on X. It is easy to see that the causal effect of Z on Y is $\beta(X)\tau(X)$, which is equal to the Intent-to-Treat and that dividing by $\beta(X)$ gives $\tau(X) = (E[Y|X, Z = 1] - E[Y|X, Z = 0])/(E[D|X, Z = 1] - E[D|X, Z = 0])$. As a consequence, we have $\tau(X) = E[Y_1 - Y_0|X, D_1 > D_0]$, the LATE conditional on X. Finally we have $\mu(X) = E[Y_0|X] + P[D_0 = 1](E[Y_1 - Y_0|X, D_0 = 1] - E[Y_1 - Y_0|X, D_1 > D_0]) \neq \mu_0(X)$. Suppose that we have a ML estimate of the LATE conditional on X that we denote $\widehat{\tau}(X)$ – the so-called "proxy predictor". Notice that $\widehat{\tau}(X)$ can be computed by taking the ratio of a ML proxy of the Intent-to-Treat to a ML proxy of the take-up.

Chernozhukov et al. (2018b) define three parameters of interest in that context. The first one is the Best Linear Predictor (BLP) of $\tau(X)$ using $\widehat{\tau}(X)$, *i.e.* the L^2 projection of $\tau(X)$ on a constant and $\widehat{\tau}(X)$:

$$\text{BLP}[\tau(X)|\widehat{\tau}(X)] = E[\tau(X)] + \frac{Cov(\tau(X), \widehat{\tau}(X))}{V(\widehat{\tau}(X))} \left(\widehat{\tau}(X) - E[\widehat{\tau}(X)]\right).$$

The second one is the Sorted Group Average Treatment Effect (GATES). To define it, divide the support of the proxy predictor by quantiles, defining groups that share an (estimated) treatment response in a given interval and perform inference over their expected treatment effect:

$$E[\tau(X)|G_1] \leq \ldots \leq E[\tau(X)|G_K],$$

for $G_k = \mathbf{1}\{\ell_{k-1} \leq \widehat{\tau}(X) < \ell_k\}$ with $-\infty = \ell_0 \leq \ell_1 \leq \ldots \leq \ell_K = +\infty$. In that context, G_1 is a binary variable equal to one for a lowest-responder and G_K is a binary variable equal to one for a highest-responder. The third parameter(s) of interest are

the the characteristics of the most and least affected groups, $E[X|G_1]$ and $E[X|G_K]$ for characteristics X since they are observed. This set of parameters is called Classification Analysis (CLAN).

The strategy of Chernozhukov et al. (2018b) to estimate the BLP and GATES relies on constructing an unbiased signal $\tilde{Y}$ of $\tau(X)$ and regressing it on the proxy predictor $\widehat{\tau}(X)$ or on the group membership dummies $G_1, \ldots, G_K$. In the absence of perfect compliance, because $E[\varepsilon|X, Z] \neq 0$, it is not possible to construct a signal $\tilde{Y}$ such that $E[\tilde{Y}|X] = \tau(X)$ without assuming that we have a consistent estimate of $\beta(X)$ – which in most cases is not be available, or in contradiction with the agnostic approach of assuming that $\widehat{\tau}(X)$ is merely a proxy predictor of the LATE. The CLAN, however, can still be performed.

Going further than these three types of parameters, we want to study selection into the treatment by the job counsellor and define relevant quantities to be estimated. More specifically, we are interested in the effect of the expected baseline employability $E[Y_0|X]$ and of the expected treatment effect $\tau(X)$ on the probability of being admitted into the program $\beta(X)$. An unbiased signal for the probability of being a complier is $\tilde{D}$ defined as:

$$\tilde{D} = \frac{Z - p(X)}{p(X)(1 - p(X))} D,$$

where $p(X) = P[Z = 1|X]$. It is unbiased in the sense that $E[\tilde{D}|X] = \beta(X)$. Running a regression of $\tilde{D}$ on $\widehat{\tau}(X) - E[\widehat{\tau}(X)]$ will estimate the covariance between the conditional probability of being a complier and the proxy predictor of the LATE, $Cov(\beta(X), \widehat{\tau}(X))/Var(\widehat{\tau}(X))$. The result is a straightforward adaptation of Theorem 2.1 in Chernozhukov et al. (2018b), that we state in the appendix, see Theorem 4.1. As we have highlighted above, we believe that correlation to be small in the private program. Another way to analyze the optimality of selection decisions by caseworkers, is to include the binary variable D in the CLAN to check whether high-responders are more likely to enter in the program than low-responders.

Moreover, we can also form a proxy predictor of $E[Y_0|X]$ and perform the regression of $\tilde{D}$ on that proxy predictor to gauge the correlation between the probability of being a complier and the estimated baseline probability of finding a job. Notice that the CLAN analysis based on $E[Y_0|X]$ would also be meaningful. Estimation of $E[Y_0|X]$ is not straightforward because of the selection issue. However, if there are no always-takers, *i.e.* $D_0 = 0$ a.s. we can write:

$$E[Y|Z = 0, X] = E[Y_0 + D_0(Y_1 - Y_0)|X] = E[Y_0|X] = \mu_0(X).$$

So we can estimate $E[Y_0|X]$ by using the regression function estimated over the population assigned to the standard track.

5. Results

(The empirical results in this section are preliminary.)

This section analyzes the heterogeneity in the treatment effect of both the public and the

private programs using the generic machine learning framework described in the previous section. We consider four outcomes of interest: (i) a dummy variable taking the value one if the individual has received no unemployment benefit over the first six months of the experiment, (ii) a dummy variable taking the value one if the individual has received no unemployment benefit over months 6-12, (iii) the sum of all the unemployment benefits over the first six months, (iv) the sum of all the unemployment benefits over months 6-12.

We considered four machine learning algorithms: Elastic Net, Boosting, Random Forests and Neural Networks but only report the results for the first two as they yielded the ML proxies that maximized the correlation with the true CATE, here $\beta(X)\tau(X)$. To train each algorithm, we employ a two-fold cross-validation procedure and we report the results obtained over 100 different random partitions of the data between the auxiliary sample (where the algorithms are trained, 80% of the data) and the main sample (where the treatment effect is predicted using the ML proxies obtained in the auxiliary sample, 20% of the data).

5.1. Detection of Heterogeneity

We are first interested in detecting heterogeneity in both the take up (that is $\beta(X)$) and the intent-to-treat for each of the four outcomes (*i.e.* $\beta(X)\tau(X)$). For this purpose, we estimate the BLP of each these theoretical quantities using the ML proxies obtained by an Elastic Net and a Boosting algorithm. For that, consider the theoretical quantity $S(X)$ (either $\beta(X)$ or $\beta(X)\tau(X)$) and the corresponding ML proxy $\widehat{S}(X)$. Consider the following weighted linear regression:

$$V = \beta_1(Z - p(X)) + \beta_2(Z - p(X)(\widehat{S}(X) - \mathbb{E}\widehat{S}(X)) + \varepsilon, \tag{BLP}$$

with weights given by $w(X) = [p(X)(1 - p(X))]^{-1}$. In our case, the left-hand variable V is a place-holder either for D, the entry into the treatment, or each of the four outcomes described above. Theorem 2.1 in Chernozhukov et al. (2018b) ensures that $\beta_1 = \mathbb{E}S_0(X)$ and $\beta_2 = Cov(S_0(X), \widehat{S}(X))/Var(\widehat{S}(X))$. Tables 4.2 and 4.3 report the results for β_1 and β_2 from equation (BLP) for the public program and the private program, respectively.

For both the public and the private program, assignation into the treatment has a significant impact on entry into the treatment, the take-up being estimated at 40% and 32% respectively. In both cases, there is heterogeneity in the treatment take-up, $\beta(X)$, that seems well-captured by both ML proxies as we reject the null hypothesis "$\beta_2 = 0$". For the public program, we do not detect any heterogeneity in the intent-to-treat, while only the probability of not perceived unemployment benefits over months 6 to 12 after the start of the experiment is significantly affected by the treatment. Notice that not rejecting "$\beta_2 = 0$" can be due either to the absence of heterogeneity or to the lack of performance of the ML proxies. Regarding the private program, we do not find a significant impact in term of average treatment effect. We do however, detect a lot of heterogeneity in terms of intent-to-treat for each outcome. In particular, the Elastic Net proxy is significantly correlated to the individual treatment effect for the four outcomes, while the Boosting proxy seems to capture only heterogeneity in the probability of not perceiving

Table 4.2: Public Program – BLP

	Elastic Net		Boosting	
	ATE (β_1)	HET (β_2)	ATE (β_1)	HET (β_2)
Program Entry (D)	0.408	0.298	0.408	0.402
	(0.395,0.421)	(0.158,0.436)	(0.395,0.421)	(0.115,0.600)
	[0.000]	[0.000]	[0.000]	[0.010]
No UB over 6m	0.003	-0.023	0.004	-0.026
	(-0.007,0.014)	(-0.166,0.128)	(-0.006,0.014)	(-0.154,0.099)
	[1.000]	[1.000]	[0.862]	[1.000]
No UB over 6-12m	0.021	-0.07	0.022	-0.07
	(0.004,0.038)	(-0.231,0.115)	(0.005,0.039)	(-0.236,0.122)
	[0.034]	[0.960]	[0.022]	[0.939]
Amount of UB 6m	-86.44	0.125	-80.55	0.007
	(-197.8,25.75)	(-0.134,0.358)	(-178.7,17.25)	(-0.057,0.076)
	[0.260]	[0.779]	[0.214]	[1.000]
Amount of UB 6-12m	-139.6	0.031	-124.9	0.002
	(-348.7,66.40)	(-0.059,0.125)	(-317.4,66.33)	(-0.063,0.069)
	[0.370]	[1.000]	[0.405]	[1.000]

Note: This table reports point estimates, .95 confidence interval and p-value for the test of nullity of each coefficient from the weighted linear regression (BLP) for the public program.

any unemployment benefit.

We do not find exactly the same results regarding the average treatment effects as Behaghel et al. (2014), who originally found that participation in the public program increased the chances of returning to employment after 6 months by 9.1 pp compared to the baseline exit rate and that participation in the private program increased the chances of returning to employment after 6 months by 4.2 pp. In order to compare these results with our ITT estimates (β_1) displayed in Tables 4.2 and 4.3, we multiply LATE estimates from Behaghel et al. (2014) cited above with the average take-up rate for each program, which leads to an ITT estimate of approximately 3.5 pp for the public program and of about 1.9 pp for the private program, a magnitude quite different from what is shown in Tables 4.2 and 4.3. Two reasons explain this discrepancy. First, we do not consider the same outcome variables, ours are based on post-treatment unemployment benefits, while the original paper had results directly on job findings. These outcomes are interesting when considering the budgetary constraint of the government but not receiving any unemployment benefits during a certain period can only be considered a proxy for return to employment – keep in mind that we restrict our sample to individuals who are eligible to benefits. Second, we consider a slightly different sample from the one used by Behaghel et al. (2014): our dataset contains about 10,000 fewer jobseekers from the inflow eligible to unemployment benefits, which may result in a sufficient loss of power to be unable to detect the reported effect in the original study. We use this sample due to the wealth of information it provides, especially regarding past unemployment benefits, number of unemployment days, and number of part-time work hours up to 5 years before treatment

Table 4.3: Private Program – BLP

	Boosting		Elastic Net	
	ATE (β_1)	HET (β_2)	ATE (β_1)	HET (β_2)
Program Entry (D)	0.325	0.809	0.326	0.675
	(0.309,0.342)	(0.706,0.913)	(0.309,0.342)	(0.579,0.776)
	[0.000]	[0.000]	[0.000]	[0.000]
No UB over 6m	-0.002	1.026	0.000	0.327
	(-0.012,0.008)	(0.930,1.122)	(-0.010,0.009)	(0.267,0.378)
	[1.000]	[0.000]	[1.000]	[0.000]
No UB over 6-12m	-0.007	0.288	-0.007	0.137
	(-0.025,0.010)	(0.138,0.438)	(-0.025,0.010)	(0.032,0.246)
	[0.827]	[0.000]	[0.834]	[0.016]
Amount of UB 6	-6.832	0.056	-2.203	0.134
	(-141.2,127.4)	(-0.304,0.398)	(-117.8,113.3)	(0.067,0.206)
	[1.000]	[1.000]	[1.000]	[0.000]
Amount of UB 6-12m	-154.20	-0.338	-83.67	0.095
	(-411.6,105.2)	(-0.546,-0.067)	(-316.2,146.2)	(0.029,0.163)
	[0.496]	[0.013]	[0.949]	[0.012]

Note: This table reports point estimates, .95 confidence interval and p-value for the test of nullity of each coefficient from the weighted linear regression (BLP) for the private program.

(and up to 10 years after the beginning of the treatment). We considered these covariates as additional and relevant data to train machine learning models.

5.2. Dimension of Heterogeneity (CLAN)

Since we detected some heterogeneity in both the take-up and the intent-to-treat, we perform the CLAN. If the dimensions of heterogeneity differs between $\beta(X)$ and $E[Y|X, Z = 1] - E[Y|X, Z = 0]$, then we can conclude that there is some heterogeneity in the conditional LATE which is the ratio of these two quantities. Table 4.4 reports the results for the public program and 4.5 reports the results for the private program, for the probability of not receiving any unemployment benefits during months 6-12.

We see that the composition of groups defined in terms of quantiles of take-up and ITT differ along a number of characteristics, suggesting heterogeneity in the conditional LATE. In particular, for the public program, there are more individuals targeting manager-level positions amongst the least likely to enter, although they are not particularly over-represented amongst the "low-responders" in terms of ITT. The unskilled and high-school dropouts are over-represented among the least likely to enter, but also among the lowest-responders. Conversely, people with more than 5 years of working experiences are over-represented among the most likely to enter and among the highest-responders. We find similar evidence for the private program although the dimensions are not the same. This time unskilled blue collars are not particularly over or under represented in the least or most likely to enter.

Table 4.4: Dimensions of Heterogeneity (CLAN) in Take-Up and ITT, Public Program

	Selection, $\beta(X)$			**Intent-To-Treat,** $E[Y\|X,Z=1]-E[Y\|X,Z=0]$		
	Most likely to Enter	Least likely to Enter	Difference	Most Affected	Least Affected	Difference
Manager/Engineer	0.064	0.103	-0.033	0.095	0.083	0.014
	(0.053,0.075)	(0.092,0.114)	(-0.049,-0.018)	(0.084,0.107)	(0.072,0.094)	(-0.001,0.029)
	-	-	[0.000]	-	-	[0.150]
Unskilled blue collar	0.073	0.094	-0.026	0.068	0.095	-0.029
	(0.062,0.084)	(0.083,0.104)	(-0.042,-0.010)	(0.057,0.079)	(0.085,0.106)	(-0.044,-0.014)
	-	-	[0.003]	-	-	[0.000]
Unskilled employee	0.100	0.189	-0.089	0.131	0.160	-0.021
	(0.086,0.114)	(0.176,0.203)	(-0.108,-0.070)	(0.117,0.144)	(0.146,0.174)	(-0.041,-0.002)
	-	-	[0.000]	-	-	[0.105]
French citizen	0.945	0.799	0.151	0.881	0.869	0.011
	(0.932,0.958)	(0.786,0.812)	(0.133,0.168)	(0.868,0.895)	(0.855,0.882)	(-0.007,0.030)
	-	-	[0.000]	-	-	[0.470]
Less than high school	0.126	0.254	-0.138	0.145	0.217	-0.070
	(0.111,0.142)	(0.239,0.270)	(-0.160,-0.116)	(0.130,0.159)	(0.202,0.232)	(-0.092,-0.049)
	-	-	[0.000]	-	-	[0.000]
College degree	0.080	0.148	-0.069	0.100	0.152	-0.047
	(0.067,0.093)	(0.136,0.161)	(-0.087,-0.051)	(0.088,0.113)	(0.139,0.166)	(-0.066,-0.028)
	-	-	[0.000]	-	-	[0.000]
Aged above 56	0.008	0.084	-0.076	0.040	0.040	-0.001
	(-0.001,0.016)	(0.076,0.092)	(-0.088,-0.065)	(0.033,0.048)	(0.032,0.047)	(-0.012,0.010)
	-	-	[0.000]	-	-	[1.000]
Age below 26	0.126	0.207	-0.080	0.175	0.175	-0.001
	(0.111,0.141)	(0.192,0.221)	(-0.100,-0.059)	(0.160,0.190)	(0.160,0.190)	(-0.022,0.021)
	-	-	[0.000]	-	-	[1.000]
No experience	0.076	0.301	-0.225	0.182	0.192	-0.008
	(0.061,0.091)	(0.286,0.315)	(-0.246,-0.205)	(0.166,0.197)	(0.176,0.208)	(-0.030,0.013)
	-	-	[0.000]	-	-	[1.000]
Above 5 years of experience	0.444	0.296	0.130	0.378	0.331	0.050
	(0.426,0.463)	(0.277,0.315)	(0.103,0.156)	(0.360,0.397)	(0.312,0.350)	(0.023,0.077)
	-	-	[0.000]	-	-	[0.001]
First unemployment spell	0.121	0.109	0.005	0.078	0.141	-0.067
	(0.108,0.134)	(0.096,0.122)	(-0.012,0.023)	(0.066,0.090)	(0.129,0.153)	(-0.085,-0.049)
	-	-	[1.000]	-	-	[0.000]
Woman	0.563	0.481	0.090	0.549	0.479	0.063
	(0.544,0.583)	(0.461,0.500)	(0.062,0.117)	(0.529,0.569)	(0.459,0.498)	(0.035,0.091)
	-	-	[0.000]	-	-	[0.000]
No child	0.466	0.588	-0.122	0.530	0.562	-0.035
	(0.446,0.485)	(0.569,0.608)	(-0.150,-0.094)	(0.510,0.549)	(0.543,0.582)	(-0.063,-0.008)
	-	-	[0.000]	-	-	[0.034]
Married	0.464	0.460	0.011	0.480	0.418	0.066
	(0.445,0.484)	(0.440,0.480)	(-0.017,0.039)	(0.461,0.500)	(0.398,0.437)	(0.038,0.094)
	-	-	[0.906]	-	-	[0.000]
Economic layoff	0.098	0.096	0.010	0.107	0.075	0.029
	(0.086,0.109)	(0.084,0.108)	(-0.007,0.026)	(0.095,0.118)	(0.063,0.086)	(0.014,0.045)
	-	-	[0.598]	-	-	[0.001]

Note: This table reports the point estimate and .95 confidence interval for the average characteristics of individuals grouped by value of their ML proxy corresponding either to their expected take-up probability or to their expected ITT. The column "difference" reports the difference and p-value (between brackets) for the test of no difference between a "high" group and a "low" group. On the left side of the table, the *most likely to enter* sub-group is defined as the sub-sample of individuals among the top 20% in terms of ML proxy for the conditional take-up probability and the *less likely to enter* sub-group is the sub-sample of individuals among the bottom 20%. The right side of the table reports the same metrics for groups defined in terms of conditional ITT, where *most affected* gathers the top 20% individuals and *less affected* the bottom 20%. The outcome is the indicator of not receiving any unemployment benefits during months 6-12.

Table 4.5: Dimensions of Heterogeneity (CLAN) in Take-Up and ITT, Private Program

	Selection, $\beta(X)$			**Intent-To-Treat,** $E[Y\|X,Z=1]-E[Y\|X,Z=0]$		
	Most likely to Enter	Least likely to Enter	Difference	Most Affected	Least Affected	Difference
Manager/Engineer	0.128	0.117	0.012	0.110	0.106	0.010
	(0.120,0.137)	(0.109,0.125)	(0.001,0.024)	(0.102,0.118)	(0.099,0.113)	(-0.001,0.020)
	-	-	[0.077]	-	-	[0.183]
Unskilled blue collar	0.057	0.060	-0.004	0.034	0.090	-0.057
	(0.051,0.063)	(0.055,0.066)	(-0.012,0.004)	(0.028,0.039)	(0.084,0.095)	(-0.065,-0.049)
	-	-	[0.651]	-	-	[0.000]
Unskilled employee	0.074	0.203	-0.130	0.087	0.204	-0.114
	(0.066,0.083)	(0.195,0.212)	(-0.142,-0.118)	(0.078,0.095)	(0.195,0.213)	(-0.126,-0.102)
	-	-	[0.000]	-	-	[0.000]
French citizen	0.949	0.750	0.200	0.755	0.889	-0.136
	(0.941,0.957)	(0.741,0.758)	(0.188,0.212)	(0.746,0.764)	(0.880,0.898)	(-0.149,-0.122)
	-	-	[0.000]	-	-	[0.000]
Less than high school	0.143	0.283	-0.139	0.181	0.253	-0.079
	(0.133,0.153)	(0.273,0.293)	(-0.153,-0.124)	(0.171,0.192)	(0.243,0.263)	(-0.093,-0.064)
	-	-	[0.000]	-	-	[0.000]
College degree	0.096	0.158	-0.063	0.137	0.147	-0.009
	(0.088,0.104)	(0.150,0.166)	(-0.074,-0.051)	(0.128,0.145)	(0.138,0.156)	(-0.022,0.003)
	-	-	[0.000]	-	-	[0.281]
Aged above 56	0.000	0.245	-0.245	0.023	0.080	-0.058
	(-0.008,0.008)	(0.238,0.253)	(-0.256,-0.234)	(0.017,0.028)	(0.075,0.086)	(-0.066,-0.050)
	-	-	[0.000]	-	-	[0.000]
Age below 26	0.052	0.181	-0.128	0.161	0.128	0.036
	(0.044,0.060)	(0.172,0.189)	(-0.140,-0.117)	(0.152,0.169)	(0.119,0.137)	(0.023,0.048)
	-	-	[0.000]	-	-	[0.000]
No experience	0.015	0.294	-0.279	0.073	0.258	-0.184
	(0.007,0.024)	(0.286,0.303)	(-0.291,-0.268)	(0.064,0.082)	(0.249,0.267)	(-0.197,-0.172)
	-	-	[0.000]	-	-	[0.000]
Above 5 years of experience	0.642	0.326	0.314	0.424	0.344	0.062
	(0.631,0.654)	(0.314,0.338)	(0.297,0.331)	(0.412,0.436)	(0.332,0.356)	(0.045,0.079)
	-	-	[0.000]	-	-	[0.000]
First unemployment spell	0.123	0.175	-0.053	0.130	0.119	0.016
	(0.114,0.132)	(0.167,0.184)	(-0.065,-0.040)	(0.122,0.138)	(0.110,0.127)	(0.005,0.027)
	-	-	[0.000]	-	-	[0.015]
Woman	0.548	0.501	0.047	0.448	0.553	-0.106
	(0.535,0.560)	(0.489,0.514)	(0.030,0.065)	(0.435,0.460)	(0.540,0.565)	(-0.123,-0.089)
	-	-	[0.000]	-	-	[0.000]
No child	0.393	0.613	-0.222	0.505	0.557	-0.043
	(0.381,0.405)	(0.601,0.625)	(-0.239,-0.205)	(0.492,0.517)	(0.545,0.569)	(-0.061,-0.026)
	-	-	[0.000]	-	-	[0.000]
Married	0.564	0.463	0.093	0.480	0.473	0.014
	(0.551,0.576)	(0.451,0.476)	(0.075,0.110)	(0.468,0.493)	(0.460,0.485)	(-0.003,0.032)
	-	-	[0.000]	-	-	[0.221]
Economic layoff	0.186	0.095	0.093	0.126	0.103	0.022
	(0.178,0.195)	(0.087,0.104)	(0.081,0.105)	(0.118,0.134)	(0.095,0.111)	(0.011,0.033)
	-	-	[0.000]	-	-	[0.000]

Note: This table reports the point estimate and .95 confidence interval for the average characteristics of individuals grouped by value of their ML proxy corresponding either to their expected take-up probability or to their expected ITT. The column "difference" reports the difference and p-value (between brackets) for the test of no difference between a "high" group and a "low" group. On the left side of the table, the *most likely to enter* sub-group is defined as the sub-sample of individuals among the top 20% in terms of ML proxy for the conditional take-up probability and the *less likely to enter* sub-group is the sub-sample of individuals among the bottom 20%. The right side of the table reports the same metrics for groups defined in terms of conditional ITT, where *most affected* gathers the top 20% individuals and *less affected* the bottom 20%. The outcome is the indicator of not receiving any unemployment benefits during months 6-12.

For entry in both programs, citizenship seems to matter a lot. French individuals represent a higher proportion of the most likely to enter than of the least likely to enter but they do not compose the most the same proportions.

Notice that when we say "over-represented" in a group we implicitly mean "relatively to the other extreme group", not to the general population.

5.3. Selection into the Treatment

To study selection into the treatment, we employ the two strategies described in the previous section: (i) a comparison of the rates of entrance intro the program for groups defined in terms of similar ML proxy levels either for the baseline outcome or for the treatment effect, (ii) a regression of the unbiased signal for the probability of being a complier, $\tilde{D} = D(Z - p(X))/(p(X)(1 - p(X)))$, on both the ML proxy for the individual baseline outcome, $\widehat{\mu}_0(X)$, and the ML proxy for the individual treatment effect, $\widehat{\tau}(X)$. Tables 4.6 and 4.7 report the results for the public program, while Tables 4.8 and 4.9 report the results for the private program.

In the case of the public program, these results have to be interpreted carefully as we did not detect any treatment effect heterogeneity except for one outcome. We can however interpret results regarding the baseline outcomes. From Table 4.6, it seems that individuals that are classified among the top 20% in terms of probability of not receiving unemployment benefits over the first six months without treatment by both the Boosting and the Elastic Net proxies enter into the treatment at a significantly lower rate than those who are classified amongst the bottom 20%. Conversely, individuals who are classified among the top 20% in terms of their baseline amount of unemployment benefits over the first six months enter the treatment at a significantly higher rate that those who are classified amongst the bottom 20%. Table 4.7 provides qualitatively similar results where a 1 pp increase in the probability of not receiving any unemployment benefits over the first six months roughly translates into a decrease of .3 pp in the probability of enrollment. This suggests that case-workers in the public program targeted people with fewer labor market prospects.

Regarding the private program, the results are qualitatively similar albeit clearer: individuals who appear to be more at risk of unemployment are accepted into the program at significantly higher rates. There is a about a 6% difference when groups are defined in terms of probability of not perceiving unemployment benefits, and a 12% difference when groups are defined in terms of amount of unemployment benefits. Although we detected heterogeneity in the treatment effect, individual expected treatment effect does not appear to affect enrollment rates.

6. Conclusion

In this chapter, we revisited a randomized experiment designed to compare public and private provision of job counselling using machine learning tools. The difference in the payment structure between the two arms of the treatment made it likely that job counsellors had different incentives to select randomized applicants. In particular, we studied

Table 4.6: Selection into the Public Program – Group Comparison

$\mu_0(X)$	Elastic Net			Boosting		
	High Baseline	Low Baseline	Difference	High Baseline	Low Baseline	Difference
No UB over 6m	0.209	0.246	-0.038	0.192	0.221	-0.029
	(0.191,0.226)	(0.229,0.263)	(-0.061,-0.014)	(0.174,0.210)	(0.198,0.245)	(-0.056,-0.001)
	-	-	[0.003]	-	-	[0.092]
No UB over 6-12m	0.218	0.230	-0.011	0.212	0.236	-0.027
	(0.200,0.236)	(0.214,0.246)	(-0.035,0.012)	(0.195,0.229)	(0.219,0.254)	(-0.052,-0.001)
	-	-	[0.702]	-	-	[0.088]
Amount of UB 6m	0.214	0.160	0.049	0.240	0.213	0.027
	(0.192,0.237)	(0.139,0.181)	(0.017,0.082)	(0.223,0.256)	(0.197,0.230)	(0.003,0.050)
	-	-	[0.006]	-	-	[0.054]
Amount of UB 6-12m	0.233	0.208	0.023	0.235	0.217	0.018
	(0.215,0.250)	(0.191,0.225)	(-0.002,0.049)	(0.218,0.251)	(0.201,0.234)	(-0.006,0.041)
	-	-	[0.145]	-	-	[0.285]
$\tau(X)$	Elastic Net			Boosting		
	Most Affected	Least Affected	Difference	Most Affected	Least Affected	Difference
No UB over 6m	0.218	0.219	-0.001	0.218	0.231	-0.011
	(0.202,0.234)	(0.202,0.235)	(-0.024,0.022)	(0.201,0.234)	(0.215,0.247)	(-0.033,0.011)
	-	-	[1.000]	-	-	[0.730]
No UB over 6-12m	0.224	0.216	0.010	0.229	0.223	0.004
	(0.208,0.241)	(0.200,0.233)	(-0.013,0.033)	(0.213,0.246)	(0.207,0.240)	(-0.019,0.028)
	-	-	[0.788]	-	-	[1.000]
Amount of UB 6m	0.199	0.245	-0.048	0.223	0.234	-0.013
	(0.183,0.215)	(0.229,0.262)	(-0.071,-0.024)	(0.206,0.239)	(0.217,0.250)	(-0.036,0.011)
	-	-	[0.000]	-	-	[0.573]
Amount of UB 12m	0.216	0.213	0.002	0.222	0.232	-0.010
	(0.200,0.232)	(0.197,0.229)	(-0.021,0.025)	(0.206,0.239)	(0.216,0.249)	(-0.033,0.013)
	-	-	[1.000]	-	-	[0.793]

Note: This table reports the point estimate and .95 confidence interval for the probability of entering into the public program for individuals grouped by value of their ML proxy corresponding either to their expected outcome without treatment or to their expected LATE. The column "difference" reports the difference and p-value (between brackets) for the test of no difference between a "high" group and a "low" group. The upper-half compute these quantities for sub-groups defined in terms of estimated baseline level of the outcome, with the outcome varying at each row. The *high baseline* is defined as the sub-sample of individuals among the top 20% in terms of ML proxy for the baseline outcome and the *low baseline* is the sub-sample of individuals among the bottom 20%. The lower-half reports the same metrics for groups defined in terms of conditional LATE.

Table 4.7: Selection into the Public Program – Regression

	Elastic Net		Boosting	
	Baseline Outcome $\mu_0(X)$	Expected LATE $\tau(X)$	Baseline Outcome $\mu_0(X)$	Expected LATE $\tau(X)$
No UB over 6m	-0.280	0.000	-0.165	-0.063
	(-0.486,-0.097)	(-0.002,0.008)	(-0.354,0.051)	(-0.184,0.056)
	[0.006]	[1.000]	[0.278]	[0.565]
No UB over 6-12m	-0.326	0.000	-0.383	0.000
	(-0.547,-0.090)	(-0.030,0.003)	(-0.723,-0.061)	(-0.108,0.112)
	[0.012]	[1.000]	[0.039]	[1.000]
Amount of UB 6m	0.000	0.000	0.000	0.000
	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)
	[0.792]	[0.971]	[0.001]	[0.947]
Amount of UB 6-12m	0.000	0.000	0.000	0.000
	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)
	[0.812]	[1.000]	[0.003]	[0.823]

Note: This table reports results from the regression of an unbiased signal of entry intro the program, $\tilde{D} = D(Z - p(X))/(p(X)(1 - p(X)))$, on the ML proxies for the individual baseline outcome and the individual treatment effect.

Table 4.8: Selection into the Private Program – Group Comparison

$\mu_0(X)$	Elastic Net			Boosting		
	High Baseline	Low Baseline	Difference	High Baseline	Low Baseline	Difference
No UB over 6m	0.325	0.385	-0.059	0.320	0.378	-0.055
	(0.314,0.337)	(0.373,0.396)	(-0.076,-0.042)	(0.308,0.331)	(0.366,0.390)	(-0.072,-0.038)
	-	-	[0.000]	-	-	[0.000]
No UB over 6-12m	0.292	0.360	-0.069	0.290	0.358	-0.069
	(0.280,0.303)	(0.348,0.371)	(-0.085,-0.053)	(0.279,0.302)	(0.346,0.369)	(-0.085,-0.052)
	-	-	[0.000]	-	-	[0.000]
Amount of UB 6m	0.297	0.321	-0.022	0.381	0.263	0.117
	(0.280,0.314)	(0.308,0.334)	(-0.045,0.001)	(0.370,0.393)	(0.251,0.274)	(0.101,0.133)
	-	-	[0.118]	-	-	[0.000]
Amount of UB 6-12m	0.408	0.327	0.087	0.378	0.253	0.122
	(0.396,0.420)	(0.315,0.339)	(0.070,0.105)	(0.367,0.389)	(0.242,0.265)	(0.105,0.138)
	-	-	[0.000]	-	-	[0.000]
$\tau(X)$	Elastic Net			Boosting		
	Most Affected	Least Affected	Difference	Most Affected	Least Affected	Difference
No UB over 6m	0.326	0.269	0.061	0.324	0.294	0.030
	(0.315,0.338)	(0.258,0.281)	(0.044,0.077)	(0.313,0.336)	(0.282,0.305)	(0.014,0.046)
	-	-	[0.000]	-	-	[0.001]
No UB over 6-12m	0.323	0.274	0.047	0.323	0.287	0.039
	(0.311,0.334)	(0.263,0.285)	(0.031,0.063)	(0.312,0.335)	(0.275,0.298)	(0.023,0.055)
	-	-	[0.000]	-	-	[0.000]
Amount of UB 6m	0.297	0.259	0.029	0.301	0.340	-0.044
	(0.286,0.308)	(0.248,0.270)	(0.013,0.045)	(0.289,0.312)	(0.328,0.351)	(-0.060,-0.027)
	-	-	[0.001]	-	-	[0.000]
Amount of UB 12m	0.288	0.308	-0.024	0.297	0.343	-0.043
	(0.277,0.300)	(0.297,0.319)	(-0.040,-0.008)	(0.286,0.309)	(0.331,0.354)	(-0.060,-0.027)
	-	-	[0.009]	-	-	[0.000]

Note: This table reports the point estimate and .95 confidence interval for the probability of entering into the private program for individuals grouped by value of their ML proxy corresponding either to their expected outcome without treatment or to their expected LATE. The column "difference" reports the difference and p-value (between brackets) for the test of no difference between a "high" group and a "low" group. The upper-half compute these quantities for sub-groups defined in terms of estimated baseline level of the outcome, with the outcome varying at each row. The *high baseline* is defined as the sub-sample of individuals among the top 20% in terms of ML proxy for the baseline outcome and the *low baseline* is the sub-sample of individuals among the bottom 20%. The lower-half reports the same metrics for groups defined in terms of conditional LATE.

Table 4.9: Selection into the Private Program – Regression

	Elastic Net		Boosting	
	Baseline Outcome $\mu_0(X)$	Expected LATE $\tau(X)$	Baseline Outcome $\mu_0(X)$	Expected LATE $\tau(X)$
No UB over 6m	-0.053	0.000	-0.192	0.000
	(-0.141,0.033)	(0.000,0.000)	(-0.266,-0.120)	(0.000,0.000)
	[0.435]	[1.000]	[0.000]	[1.000]
No UB over 6-12m	-0.164	0.000	-0.218	0.000
	(-0.254,-0.075)	(0.000,0.000)	(-0.295,-0.140)	(0.000,0.000)
	[0.001]	[1.000]	[0.000]	[1.000]
Amount of UB 6m	0.000	0.000	0.000	0.000
	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)
	[0.012]	[1.000]	[0.000]	[1.000]
Amount of UB 6-12m	0.000	0.000	0.000	0.000
	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)
	[1.000]	[1.000]	[0.000]	[1.000]

Note: This table reports results from the regression of an unbiased signal of entry intro the program, $\tilde{D} = D(Z - p(X))/(p(X)(1 - p(X)))$, on the ML proxies for the individual baseline outcome and the individual treatment effect.

how private providers had the incentive to either cream-skim or park candidates. Using the generic machine learning methodology developed in Chernozhukov et al. (2018b), we are able to find evidence of heterogeneous treatment effects and construct ML proxies for the individualized treatment effect. We then study the impact of both the individualized treatment effect and the individualized baseline probability of finding a job on the likelihood of enrollment in both programs.

Very preliminary results show that contrary to our assumptions, both public and private providers had a tendency to target individuals with gloomier baseline labor market prospects, while we do not find any evidence of selection based on the expected individual treatment effect.

From a methodological standpoint, we have discussed the extent to which the generic machine learning framework can be applied to randomized experiments with imperfect compliance.

7. Appendix: Descriptive Statistics

Table 4.10: Descriptive characteristics among experimental groups

	Standard	Public	(1)-(2)	Private	(1)-(3)
Age below 26	0.17	0.17		0.16	
Aged above 56	0.05	0.05		0.05	
College degree	0.14	0.14		0.14	
Less than high school	0.20	0.20		0.21	***
No experience	0.15	0.15		0.14	**
Above 5 years of experience	0.39	0.40		0.40	***
Woman	0.52	0.51		0.50	**
Married	0.46	0.45	*	0.47	
French citizen	0.86	0.85		0.84	***
No child	0.54	0.56	*	0.54	
First unemployment spell	0.13	0.12	***	0.13	
Manager/Engineer	0.11	0.11		0.12	*
Unskilled employee	0.14	0.14		0.14	
Unskilled blue collar	0.06	0.06		0.06	***
Economic layoff	0.12	0.12		0.12	**

Note: Each cell displays a proportion. Columns (1), (2) and (4) characterize jobseekers by their random assignment; columns (1)-(2) and (1)-(4) report the significance level of the difference between the coefficients in stated columns. *, **, ***: significance at 10%,5% and 1%. Observations are weighted by the inverse of the assignment probability.

8. Appendix: Adaptation of Th. 2.1 in Chernozhukov et al. (2018b)

Suppose that we are interested in an unobservable quantity $S_0(X)$ and observe the random variable V. Suppose also that we can construct an unbiased signal of $S_0(X)$, *i.e.* the random variable $w(X)(D - p(X))V$ with $w(X) = [p(X)(1 - p(X))]^{-1}$ that has the property $E[w(X)(Z - p(X))V|X] = S_0(X)$. We also assume $V = b_0(X) + ZS_0(X) + U$, where $E[U|X, Z] = 0$. On the other hand, we have a ML proxy $x \to \widehat{m}(x)$, obtained on the auxiliary sample, that may or may not be an estimate of $S_0(X)$ but for which we want to learn the best linear predictor of $S_0(X)$ using $\widehat{m}(X)$, *i.e.* the L^2 projection of $S_0(X)$ on a constant and $\widehat{m}(X)$:

$$\text{BLP}[S_0(X)|\widehat{m}(X)] = E[S_0(X)] + \frac{Cov(S_0(X), \widehat{m}(X))}{V(\widehat{m}(X))} (\widehat{m}(X) - E[\widehat{m}(X)]) .$$

To obtain an estimate of $Cov(S_0(X), \widehat{m}(X))/V(\widehat{m}(X))$ we can regress the unbiased signal $w(X)(D - p(X))V$ on $\widehat{m}(X) - E[\widehat{m}(X)]$ in the main sample, that is, run the regression:

$$w(X)(Z - p(X))V = \beta_1 + \beta_2(\widehat{m}(X) - E[\widehat{m}(X)]) + \varepsilon,\ E[\varepsilon(1, (\widehat{m}(X) - E[\widehat{m}(X)]))'] = 0. \tag{4.1}$$

Theorem 4.1 (Adaptation of Th. 2.1 in Chernozhukov et al. (2018b)) *Consider $x \to \widehat{m}(x)$ as a fixed map. Assume that V has finite second moment and that $V(\widehat{m}(X)) \neq 0$. Then (β_1, β_2) defined in (4.1) are the coefficient of the BLP of $S_0(x)$ given $\widehat{m}(X)$:*

$$\beta_1 = E[S_0(X)] \text{ and } \beta_2 = \frac{Cov(S_0(X), \widehat{m}(X))}{V(\widehat{m}(X))}.$$

Proof of Theorem 4.1

We only show that $\beta_2 = Cov(S_0(X), \widehat{m}(X))/V(\widehat{m}(X))$ since the proof for β_1 is similar. The normal equations that define (β_1, β_2) in Equation 4.1 give for β_2:

$$\beta_2 = \frac{Cov(w(X)(Z - p(X))V, \widehat{m}(X) - E[\widehat{m}(X)])}{V(\widehat{m}(X) - E[\widehat{m}(X)])}.$$

The denominator is equal to $V(\widehat{m}(X))$. Now since $\widehat{m}(X) - E[\widehat{m}(X)]$ has mean zero, the numerator is:

$$Cov(w(X)(Z - p(X))V, \widehat{m}(X) - E[\widehat{m}(X)]) = E[w(X)(Z - p(X))V(\widehat{m}(X) - E[\widehat{m}(X)])].$$

Notice that $E[w(X)(Z - p(X))Z|X] = [w(X)(Z - p(X))^2|X] = 1$ since $Z|X \sim \mathcal{B}(p(X))$. Recall that $V = b_0(X) + ZS_0(X) + U$, the law of iterated expectations yields:

$$E[w(X)(Z - p(X))b_0(X)(\widehat{m}(X) - E[\widehat{m}(X)])] = E[w(X)b_0(X)(\widehat{m}(X) - E[\widehat{m}(X)]) \underbrace{E[Z - p(X)|X]}_{=0}] = 0$$

$$E[w(X)(Z - p(X))ZS_0(X)(\widehat{m}(X) - E[\widehat{m}(X)])] = E[S_0(X)(\widehat{m}(X) - E[\widehat{m}(X)])] = Cov(S_0(X), \widehat{m}(X)),$$
$$E[w(X)(Z - p(X))U(\widehat{m}(X) - E[\widehat{m}(X)])] = E[w(X)(Z - p(X)) \underbrace{E[U|X, Z]}_{=0}(\widehat{m}(X) - E[\widehat{m}(X)])] = 0,$$

proving that $\beta_2 = Cov(S_0(X), \widehat{m}(X))/V(\widehat{m}(X))$. □

In the main text, we use this strategy for several instances of $S_0(X)$ and $\widehat{m}(X)$. In particular, we consider $S_0(X) = \beta(X) = P[D_1 > D_0|X]$ with the signal $V = D$ for entry into the program. While we formed $\widehat{m}(X) = \widehat{\tau}(X)$ a ML proxy for the conditional LATE.

Bibliography

Aaron, C., Cholaquidis, A., and Fraiman, R. (2017). A generalization of the maximal-spacings in several dimensions and a convexity test. *Extremes*, 20(3):605–634.

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231 – 263.

Abadie, A. (2019). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Working Paper*.

Abadie, A. and Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics*, 10(1):465–503.

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510.

Abadie, A. and Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, 93(1):113–132.

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11.

Abadie, A. and Imbens, G. W. (2012). A martingale representation for matching estimators. *Journal of the American Statistical Association*, 107(498):833–843.

Abadie, A. and Kasy, M. (2018). Choosing among regularized estimators in empirical economics: The risk of machine learning. *The Review of Economics and Statistics*, 0(ja):null.

Abadie, A. and Spiess, J. (2016). Robust post-matching inference. *Working Paper*.

Acemoglu, D., Johnson, S., Kermani, A., Kwak, J., and Mitton, T. (2016). The value of connections in turbulent times: Evidence from the united states. *Journal of Financial Economics*, 121:368–391.

Addison, J. T., Blackburn, M. L., and Cotti, C. D. (2014). On the Robustness of Minimum Wage Effects: Geographically-Disparate Trends and Job Growth Equations. Working Paper Series in Economics 330, University of Lüneburg, Institute of Economics.

Amjad, M., Shah, D., and Shen, D. (2018). Robust synthetic control. *Journal of Machine Learning Research*, 19(22):1–51.

Angrist, J. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 1 edition.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2018). Synthetic Difference in Differences. *arXiv e-prints*, page arXiv:1812.09970.

Athey, S. (2015). Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 5–6, New York, NY, USA. ACM.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2017). Matrix Completion Methods for Causal Panel Data Models. *arXiv e-prints*, page arXiv:1710.10251.

Athey, S., Bayati, M., Imbens, G., and Qu, Z. (2019). Ensemble methods for causal effects in panel data settings. *AEA Papers and Proceedings*, 109:65–70.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Athey, S. and Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11(1):null.

Athey, S. and Wager, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Bach, P., Chernozhukov, V., and Spindler, M. (2018). Closing the U.S. gender wage gap requires understanding its heterogeneity. *arXiv e-prints*, page arXiv:1812.04345.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Behaghel, L., Crépon, B., and Gurgand, M. (2014). Private and public provision of counseling to job seekers: Evidence from a large controlled experiment. *American Economic Journal: Applied Economics*, 6(4):142–74.

Behaghel, L., Crépon, B., and Gurgand, M. (2012). Private and public provision of counseling to job-seekers: Evidence from a large controlled experiment. IZA Discussion Papers 6518, Institute of Labor Economics (IZA).

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014a). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2):29–50.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014b). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Belloni, A., Chernozhukov, V., Hansen, C., and Kozbur, D. (2016). Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics*, 34(4):590–605.

Ben-Michael, E., Feller, A., and Rothstein, J. (2019). The Augmented Synthetic Control Method. *arXiv e-prints*, page arXiv:1811.04170.

Berg, M. d., Cheong, O., Kreveld, M. v., and Overmars, M. (2008). *Computational Geometry: Algorithms and Applications*. Springer-Verlag TELOS, Santa Clara, CA, USA, 3rd ed. edition.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732.

Bilgel, F. and Galle, B. (2015). Financial incentives for kidney donation: A comparative case study using synthetic controls. *Journal of Health Economics*, 43(C):103–117.

Bohn, S., Lofstrom, M., and Raphael, S. (2014a). Did the 2007 Legal Arizona Workers Act reduce the state's unauthorized immigrant population? *Review of Economics and Statistics*, 96(2):258–269.

Bohn, S., Lofstrom, M., and Raphael, S. (2014b). Did the 2007 legal arizona workers act reduce the state's unauthorized immigrant population? *The Review of Economics and Statistics*, 96(2):258–269.

Boissonnat, J.-D., Devillers, O., and Hornus, S. (2009). Incremental construction of the Delaunay graph in medium dimension. In *Annual Symposium on Computational Geometry*, pages 208–216, Aarhus, Denmark.

Boissonnat, J.-D. and Yvinec, M. (1998). *Algorithmic Geometry*. Cambridge University Press, New York, NY, USA.

Brunel, V.-E. (2017). Uniform deviation and moment inequalities for random polytopes with general densities in arbitrary convex bodies. *arXiv e-prints*, page arXiv:1704.01620.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018a). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2018b). Generic machine learning inference on heterogenous treatment effects in randomized experiments. Working Paper 24678, National Bureau of Economic Research.

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268.

Chernozhukov, V., Goldman, M., Semenova, V., and Taddy, M. (2017a). Orthogonal Machine Learning for Demand Estimation: High Dimensional Causal Inference in Dynamic Panels. *arXiv e-prints*, page arXiv:1712.09988.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments. *American Economic Review*, 105(5):486–90.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annu. Rev. Econ.*, 7(1):649–688.

Chernozhukov, V., Nekipelov, D., Semenova, V., and Syrgkanis, V. (2018). Plug-in Regularized Estimation of High-Dimensional Parameters in Nonlinear Semiparametric Models. *arXiv e-prints*, page arXiv:1806.04823.

Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2017b). An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls. *arXiv e-prints*, page arXiv:1712.09089.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98(464):900–916.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2008). Nonparametric tests for treament effect heterogeneity. *Review of Economics and Statistics*, 90(3):389–405.

Cunningham, S. and Shah, M. (2018). Decriminalizing indoor prostitution: Implications for sexual violence and public health. *The Review of Economic Studies*, 85(3):1683–1715.

Davis, J. M. and Heller, S. B. (2017a). Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs. Working Paper 23443, National Bureau of Economic Research.

Davis, J. M. and Heller, S. B. (2017b). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107(5):546–50.

Dehejia, R. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84(1):151–161.

Deville, J.-C., Sarndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88(423):1013–1020.

Devillers, O. and Teillaud, M. (2003). Perturbations and Vertex Removal in a 3D Delaunay Triangulation. In *14th ACM-Siam Symposium on Discrete Algorithms (SODA)*, pages 313–319, Baltimore, MA, United States.

Dietrichson, J. and Ellegård, L. M. (2015). Assist or desist? Conditional bailouts and fiscal discipline in local governments. *European Journal of Political Economy*, 38(C):153–168.

Doudchenko, N. and Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. *NBER Working Papers*, 22791.

Dube, A. and Zipperer, B. (2015). Pooling Multiple Case Studies Using Synthetic Controls: An Application to Minimum Wage Policies. IZA Discussion Papers 8944, Institute for the Study of Labor (IZA).

Efron, B. (1965). Increasing properties of polya frequency function. *Ann. Math. Statist.*, 36(1):272–279.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1 – 23.

Fermand, B. and Pinto, C. (2019). Synthetic control with imperfect treatment fit. *Working Paper*.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1):259–276.

Firpo, S. and Possebom, V. (2018). Synthetic Control Method: Inference, Sensitivity Analysis and Confidence Sets. *Journal of Causal Inference*, 6(2):1–26.

Firpo, S. and Ridder, G. (2008). Bounds on functionals of the distribution of treatment effects. *Working Paper*.

Gaillac, C. and L'Hour, J. (2019). Machine learning for econometrics, lecture notes ensae paris.

Gobillon, L. and Magnac, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *The Review of Economics and Statistics*, 98(3):535–551.

Graham, B. S., Pinto, C. C. D. X., and Egel, D. (2012). Inverse Probability Tilting for Moment Condition Models with Missing Data. *Review of Economic Studies*, 79(3):1053–1079.

Hackmann, M. B., Kolstad, J. T., and Kowalski, A. E. (2015). Adverse selection and an individual mandate: When theory meets practice. *American Economic Review*, 105(3):1030–1066.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46.

Hall, P. and Heyde, C. (1980). *Martingale limit theory and its application*. Probability and mathematical statistics. Academic Press.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263.

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Number 9780521885881 in Cambridge Books. Cambridge University Press.

Janson, S. (1987). Maximal spacings in several dimensions. *Ann. Probab.*, 15(1):274–280.

Joag-Dev, K. and Proschan, F. (1983). Negative association of random variables with applications. *Ann. Statist.*, 11(1):286–295.

Kitagawa, T. and Muris, C. (2016). Model averaging in semiparametric estimation of treatment effects. *Journal of Econometrics*, 193(1):271 – 289.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics*, 133(1):237–293.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Sunstein, C. R. (2019). Discrimination in the age of algorithms. Working Paper 25548, National Bureau of Economic Research.

Kleven, H. J., Landais, C., and Saez, E. (2013). Taxation and international migration of superstars: Evidence from the European football market. *American Economic Review*, 103(5):1892–1924.

Kline, P. (2011). Oaxaca-blinder as a reweighting estimator. *American Economic Review*, 101(3):532–37.

Klößner, S., Kaul, A., Pfeifer, G., and Schieler, M. (2018). Comparative politics and the synthetic control method revisited: a note on Abadie et al. (2015). *Swiss Journal of Economics and Statistics*, 154(1):1–11.

Knaus, M. C., Lechner, M., and Strittmatter, A. (2018). Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence. IZA Discussion Papers 10961, Institute for the Study of Labor (IZA).

Kreif, N., Grieve, R., Hangartner, D., Turner, A. J., Nikolova, S., and Sutton, M. (2016). Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health Economics*, 25(12):1514–1528.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165.

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review*, 76(4):604–20.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, 73(1):31–43.

Leeb, H. (2006). *The distribution of a linear predictor after model selection: Unconditional finite-sample distributions and asymptotic approximations*, volume Number 49 of *Lecture Notes–Monograph Series*, pages 291–311. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, null:21–59.

Leeb, H. and Pötscher, B. M. (2008a). Recent developments in model selection and related areas. *Econometric Theory*, 24:319–322.

Leeb, H. and Pötscher, B. M. (2008b). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics*, 142(1):201–211.

Moller, J. (1994). *Lectures on Random Voronoi Tessellations*, volume 87. Springer-Verlag New York.

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Nannicini, T. and Billmeier, A. (2011). Economies in Transition: How Important Is Trade Openness for Growth? *Oxford Bulletin of Economics and Statistics*, 73(3):287–314.

Newey, W. K. and McFadden, D. (1994). Chapter 36 large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4 of *Handbook of Econometrics*, pages 2111 – 2245. Elsevier.

Okabe, A., Boots, B., Sugihara, K., and Chiu, S. N. (2000). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. Series in Probability and Statistics. John Wiley and Sons, Inc.

Oliveira, R. I. (2016). The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3):1175–1194.

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, 29(3):61–80.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA.

Rajan, V. T. (1994). Optimality of the delaunay triangulation in $\mathbb{R}^d$. *Discrete & Computational Geometry*, 12(2):189–202.

Rosenbaum, P. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Smith, J. and Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1-2):305–353.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2):614–645.

Wassmann, P. (2015). The Economic Effect of the EU Eastern Enlargement for Border Regions in the Old Member States. Annual Conference 2015 (Muenster): Economic Development - Theory and Policy 113028, Verein für Socialpolitik / German Economic Association.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(01):57–76.

Xu, Y. and Liu, L. (2018). *gsynth: Generalized Synthetic Control Method*. R package version 1.0.9.

List of Theorems and Lemmas

1.1 Lemma (Model Selection Consistency) 5
1.2 Lemma (Density of the Post-Selection estimator, from Leeb, 2006) 5

2.1 Lemma (Balancing Weights) 23
2.1 Theorem (Double Robustness) 24
2.2 Theorem (Asymptotic Normality of the Immunized Estimator) 31
2.3 Theorem (Nuisance Parameter Estimation) 49
2.2 Lemma (A Taylor Expansion Lemma) 61

3.1 Lemma (Discrepancy Bounds) 69
3.1 Theorem (Uniqueness and Sparsity) 70
3.2 Theorem (Delaunay Property I) 71
3.3 Theorem (Delaunay Property II) 72
3.2 Lemma (Bias Bound) 74
3.4 Theorem (Consistency) 75
3.5 Theorem (Asymptotic Normality) 75
3.3 Lemma (Control of $S(\lambda)$) 76
3.4 Lemma (Optimality of Delaunay for the Compound Discrepancy, Rajan, 1994) 96
3.5 Lemma (Sum of Weights) 98
3.6 Theorem (Martingale Central Limit Theorem) 100
3.6 Lemma (Conditional Probability of a Link) 103

4.1 Theorem (Adaptation of Th. 2.1 in Chernozhukov et al. (2018b)) 128

List of Figures

1.1 Finite-sample density of $\sqrt{n}(\tilde{\tau} - \tau_0)$, $\rho = .4$ 6
1.2 Finite-sample density of $\sqrt{n}(\tilde{\tau} - \tau_0)$, $\rho = .7$ 7

2.1 Sparsity patterns of β (crosses) and μ (circles) 32
2.2 The effect of Proposition 99 on per capita tobacco consumption. 40
2.3 Cigarette consumption in California, actual and counterfactual. 41

3.1 A simple example 71
3.2 Geometric properties of penalized synthetic control estimator 73
3.3 Abnormal Returns after Geithner Announcement, non-corrected inference 90
3.4 Voter Turnout in the US and EDR Laws 92
3.5 Voter Turnout in the US and EDR Laws, by Wave of Adoption 93

List of Tables

2.1 Monte-Carlo Simulations (DGP1) 34
2.2 Monte-Carlo Simulations (DGP2: Outcome Independent from X) 35
2.3 Monte-Carlo Simulations (DGP3: Heterogeneous Treatment Effect) . . . 36
2.4 Monte-Carlo simulations (DGP4: Non-Linear Outcome Equation) 37
2.5 Average Treatment Effect on the Treated for NSW. 39

3.1 Monte-Carlo Simulations, $n_1 = 10$, $n_0 = 20$ 84
3.2 Monte-Carlo Simulations, $n_1 = 10$, $n_0 = 40$ 85
3.3 Monte-Carlo Simulations, $n_1 = 10$, $n_0 = 100$ 86
3.4 Monte-Carlo Simulations, $n_1 = 100$, $n_0 = 500$ 87
3.5 Connections to Geithner and Reactions to Treasury Secretary Announcement, Synthetic Control Inference. 89

4.1 Program participation 112
4.2 Public Program – BLP 117
4.3 Private Program – BLP 118
4.4 Dimensions of Heterogeneity (CLAN) in Take-Up and ITT, Public Program 119
4.5 Dimensions of Heterogeneity (CLAN) in Take-Up and ITT, Private Program 120
4.6 Selection into the Public Program – Group Comparison 122
4.7 Selection into the Public Program – Regression 123
4.8 Selection into the Private Program – Group Comparison 124
4.9 Selection into the Private Program – Regression 125
4.10 Descriptive characteristics among experimental groups 127

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences de l'homme et de la société (SHS)

Titre : Évaluation des politiques publiques, grande dimension et machine learning

Mots clés : Econométrie, évaluation des politiques publiques, machine learning, statistique en grande dimension, contrôle synthétique

Résumé : Cette thèse regroupe trois travaux d'économétrie liés par l'application du machine learning et de la statistique en grande dimension à l'évaluation de politiques publiques. La première partie propose une alternative paramétrique au contrôle synthétique (Abadie and Gardeazabal, 2003; Abadie et al., 2010) sous la forme d'un estimateur reposant sur une première étape de type Lasso, dont on montre qu'il est doublement robuste, asymptotiquement Normal et "immunisé" contre les erreurs de première étape. La seconde partie étudie une version pénalisée du contrôle synthétique en présence de données de nature micro-économique. La pénalisation permet d'obtenir une unité synthétique qui réalise un arbitrage entre reproduire fidèlement l'unité traitée durant la période pré-traitement et n'utiliser que des unités non-traitées suffisamment semblables à l'unité traitée. Nous étudions les propriétés de cet estimateur, proposons deux procédures de type "validation croisée" afin de choisir la pénalisation et discutons des procédures d'inférence par permutation. La dernière partie porte sur l'application du *Generic Machine Learning* (Chernozhukov et al., 2018b) afin d'étudier l'hétérogénéité des effets d'une expérience aléatoire visant à comparer la fourniture publique et privée d'aide à la recherche d'emploi. D'un point de vue méthodologique, ce projet discute l'extension du *Generic Machine Learning* à des expériences avec *compliance* imparfaite.

Title : Policy evaluation, high-dimension and machine learning

Keywords : Econometrics, policy evaluation, machine learning, high-dimensional statistics, synthetic control

Abstract : This dissertation is comprised of three essays that apply machine learning and high-dimensional statistics to causal inference. The first essay proposes a parametric alternative to the synthetic control method (Abadie and Gardeazabal, 2003; Abadie et al., 2010) that relies on a Lasso-type first-step. We show that the resulting estimator is doubly robust, asymptotically Gaussian and "immunized" against first-step selection mistakes. The second essay studies a penalized version of the synthetic control method especially useful in the presence of micro-economic data. The penalization parameter trades off pairwise matching discrepancies with respect to the characteristics of each unit in the synthetic control against matching discrepancies with respect to the characteristics of the synthetic control unit as a whole. We study the properties of the resulting estimator, propose data-driven choices of the penalization parameter and discuss randomization-based inference procedures. The last essay applies the *Generic Machine Learning* framework (Chernozhukov et al., 2018b) to study heterogeneity of the treatment in a randomized experiment designed to compare public and private provision of job counselling. From a methodological perspective, we discuss the extension of the *Generic Machine Learning* framework to experiments with imperfect compliance.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France