Instituto Tecnológico Autónomo de México

Región de confianza vía la función RWP: una aplicación a métodos de control sintéticos

Tesis

que para obtener el título de

Licenciado en Economía

presenta

Isaac Meza López

Ciudad de México. 2019

Instituto Tecnológico Autónomo de México

Región de confianza vía la función RWP: una aplicación a métodos de control sintéticos

Tesis

que para obtener el título de

Licenciado en Economía

presenta

Isaac Meza López

Asesor: Dr. Enrique Seira Bejarano

Ciudad de México. 2019

"Con fundamento en los artículos 21 y 27 de la Ley Federal del Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada **"Región de confianza vía la función RWP: una aplicación a métodos de control sintéticos"**, otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la Biblioteca Raúl Bailléres Jr., autorización para que fijen la obra en cualquier medio, incluido el electrónico, y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por tal divulgación una contraprestación".

ISAAC MEZA LÓPEZ

_____

FECHA

_____

FIRMA

Lu :

Ya sabes que la palabra trivial es la más peligrosa.

clips…clips…clips

Agradecimientos

Nuevamente me doy cuenta que ésta es la parte más difícil de escribir, y asímismo hago saber que con *suficiente cobertura* agradezco a todos.

Siempre dije que tuve mucha suerte. La primera vez que me aproxime al problema fue por una llamada muy motivante de Enrique. Me explicó que era un problema abierto el conocer la distribución asintótica del estimador SCM y me motivo a escribir un teorema. En ese momento yo nunca imaginé que podía conseguir resolver el problema. Poco después y de manera fortuita, atendí un pequeño curso de probabilidad aplicada que impartía el Profesor Blanchet. Ahí fue donde me introduje a DRO y me di cuenta que con eso inmediatamente se resolvía el problema de inferencia.

Agradezco infinitamente a Enrique - quien es un mentor para mí - y a José Blanchet, no sólo porque sin ellos no hubiera aprendido sobre DRO o sobre SCM; sino porque me han enseñado mucho, he aprendido más y me han brindado muchas atenciones. Espero poder aprender mucho más de ellos en el futuro. De la misma manera agradezco a Joyce. Como ya lo había escrito la considero como una madre en el ITAM, pero también es una gran amiga y estoy seguro que esa amistad durará toda la vida.

Le doy muchas gracias a todos mis profesores en el ITAM, en especial al Profesor Ignacio Lobato, a la Profesora Beatriz Rumbos, al Profesor César Luis y al Profesor Zeferino Parada.

Por supuesto a las personas que más agradezco son mis padres y siempre será así, pues son el apoyo más grande en mi vida. Los amo.

A mi hermanita Alcalde Alcira y a mi mejor amigo David. A los superagentes. A los ganadores del dia π, a los del club de análisis y los del club de estadística: Loko, Joaco, Juan Pablo, Loredo, Wills, Omar Pardo, Pablito, Pato, Laureano, Lalo, Lucy, Chisko, Imanol. A los del CECEC y todos de la oficinita: Berns, Saúl, Dieguito, mi amiguita Pau, Erick, Dénis, Cano, Diana(s), Diego. A Estefanía. A Andrés. A Posili.

A todos los que creyeron en mi y que constantemente me dieron motivos para seguir creciendo, aunque sus cumplidos *siempre* fueron muy generosos.

Los voy a extrañar mucho, me llevo muchos recuerdo muy bellos y me voy muy emocionado y con inmensas ganas de seguir aprendiendo.

Prefacio

En economía empírica y econometría, el investigador está interesado en evaluar el efecto de una política. El principal problema que enfrenta es cómo construir un contrafactual válido. El estándar de oro para la inferencia causal es a través de un experimento controlado aleatorio. Sin embargo, en muchos casos, los experimentos siguen siendo difíciles o imposibles de implementar, por razones financieras, políticas o éticas.

El investigador utiliza una amplia variedad de estrategias para tratar de obtener una inferencia causal a partir de datos observacionales. Un método muy popular es el *método de control sintético* (SCM, por sus siglas en inglés), que esencialmente aproxima el contrafactual de la unidad tratada como una combinación convexa de las unidades de control. Un problema abierto, y área actual de investigación es cómo realizar inferencia formal en tales estrategias de identificación. Esta tesis contribuye a esta agenda al proponer un enfoque novedoso similar a Verosimilitud Empírica (EL, por sus siglas en íngles), para recuperar las regiones de confianza de las estimaciones de interés. Además, las regiones de confianza resultantes son inmunes a las formulaciones robustas de la estrategia de identificación. La principal ventaja del procedimiento de inferencia considerado aquí, en lugar de EL, es que la definición análoga de la función de perfil no requiere que exista una verosimilitud entre un modelo alternativo verosimil P y la distribución empírica P_n. Es importante destacar que la metodología presentada aquí es lo suficientemente rica y general que se puede aplicar a otros problemas en estadística, econometría e inferencia causal; esto se puede ver como una contribución indirecta de este trabajo.

La organización de la tesis es la siguiente: El capítulo 1 motiva el problema e introduce el marco de la metodología de estimación. Los capítulos 2 - 3 presentan soluciones conocidas para obtener la distribución asintótica de los pesos para el control sintético, a partir de la cual se puede derivar la distribución asintótica del estimador del efecto del tratamiento. Asimismo, se discuten las principales desventajas de tales enfoques. El capítulo 4 presenta la metodología de inferencia basada en *optimización robusta* (DRO) y deriva los resultados principales, proporcionando la región de confianza para el estimador de SCM. Estas regiones de confianza se obtienen à la Manski, lo que significa que se considera el peor escenario para un control sintético plausible. El capítulo 5, ilustra los diferentes métodos con una aplicación empírica, retomando un artículo clásico donde se usa SCM, estableciendo formalmente la significancia de los resultados ahí presentados. Concluimos en el Capítulo 6.

Foreword

In empirical economics and econometrics, researchers are interested in evaluating the effect of a policy. The main problem they face is how to construct a valid counterfactual. The gold standard for drawing inference is through a randomized controlled experiment. However, in many cases, experiments remain difficult or impossible to implement, for financial, political, or ethical reasons.

Researchers use a wide variety of strategies for attempting to draw causal inference from observational data. A very popular method is the *synthetic control method* (SCM) which essentially approximates the counterfactual to the treated unit as a convex combination of the control units. An open problem, and current area of research is how to conduct formal inference in such identification strategies. This thesis contributes to this agenda by proposing a novel approach similar to Empirical Likelihood (EL), to recover confidence regions for the estimates of interest. Moreover, the resulting confidence regions are immune to robust formulations of the identification strategy. The main advantage of the inference procedure considered here, contrasting EL, is that the analogue definition of the profile function does not require the likelihood between an alternative plausible model P, and the empirical distribution, P_n, to exist. It is important to highlight that the methodology presented here is sufficiently rich and general that can be applied to other problems in statistics, econometrics, and causal inference - this can be viewed as a spillover contribution of this work.

The organization of the thesis is as follows: Chapter 1 motivates the problem and introduces the framework of the estimation methodology. Chapters 2-3 present known solutions to obtain the asymptotic distribution for the synthetic control weights, from which the asymptotic of the treatment effect estimator can be derived. The main disadvantages of such approaches are discussed. Chapter 4 introduces the main inference methodology based on *distributionally robust optimization* (DRO) and derives the main results - which gives the confidence region for the SCM estimator. Such confidence regions are obtained à la Manski, meaning that the worst-case scenario for a plausible synthetic control is considered. Chapter 5, illustrates the different methods with an empirical application, revisiting a classical paper where the SCM is used, establishing formally the significance of its results. We conclude in Chapter 6.

Contents

Chapter 1

Synthetic control method

Synthetic control methods (SCM) are a popular approach in causal inference in comparative studies. Essentially it constructs a weighted average of different control units as a counterfactual from where the treatment group is to be compared. Unlike difference in differences approaches, this method can account for the effects of confounders changing over time, by weighting the control group to better match the treatment group before the intervention. The main problem with this methodology is the difficulty to perform inference, this is, there is little to none knowledge in the asymptotic distribution of the SCM estimator, or its confidence interval. This is a very important problem to solve as a vast literature rest in this methodology: [ADH15; BCL+18; ADH10; AG03; PY15; BN13; AI17; CGNP13; AJK+16; RSK17].

There has also been a rich literature extending such methods: [Pow16; Xu17], nests the additive fixed effects model with synthetic controls, permitting additional flexibility to estimate causal effects in the presence of differential state level trends and shocks. To address multiplicity of solutions with dissagregated data, [AL18; DI16] 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, using regularization methods to deal with a potentially large number of possible control units. [ASS18] provides a robust generalization of the synthetic control method by de-noising the observation data via singular value thresholding - this renders the approach as robust. Further, the algorithm is extended to include regularization techniques such as ridge regression and LASSO. The paper moves beyond point estimates in establishing a Bayesian framework, which allows one to quantitatively compute the uncertainty of the results through posterior probabilities. [BMFR18] uses an outcome model to estimate the bias due to covariate imbalance and then de-biases the original SCM estimate, analogous to bias correction for inexact matching. In [CMM18] an artificial counterfactual is proposed based on a large-dimensional panel of observed time-series data from a pool of untreated peers. The methodology shares roots with the panel factor model and SCM. Finally, [Dav18] relaxes the assumption of existence of a "perfect" synthetic control, which only occurs if the outcome variable is not subject to transitory shocks, implementing a two-step approach which first generates predicted values of the outcome variables for each unit and uses these predicted values instead

of the actual values of the outcome variable when constructing the synthetic control units. This review is by no means exhaustive, but gives an illustration of the importance of the methodology in the causal inference literature in comparative case studies.

Literature tackling this problem can be divided in two approaches, (1) those work relying on the assumption that treatment units are randomly assigned and uses placebo, permutation tests, or some variant exploiting the panel data structure, to conduct inference - which are called *finite population approaches*[1] [ADH15; BCL+18; ADH10; AG03; PY15; BN13; CGNP13; AJK+16; RSK17; Xu17; AL18; DI16; BMFR18; FP17; SV18; HS17; CWZ17], and (2) asymptotic approaches [WHI+15; CMM18; Pow16; Li17], where the key assumptions makes the number of individuals or time periods tend to infinity. This literature often focus on testing hypotheses about average effects over time and require the number of pre-period and post-treatment periods to tend to infinity.

The main disadvantage with the first approach is that the graphical analysis with placebos can be misleading, as placebo runs with lower expected squared prediction errors would still be considered in the analysis. [HS17] address a setting where permutation tests may be distorted. The validity of such tests requires a strong normality distribution assumption for the idiosyncratic error under a factor model data generating framework. Moreover, inference in such models is complicated by the fact that errors might exhibit intra-group and serial correlations (few treated groups and heteroskedastic errors). [CWZ17] approach will instead carry out the permutations over stochastic errors in the potential outcomes with respect to time, and not the cross-sectional units. These types of permutations rely on weak dependence of stochastic errors over time rather than exchangeability across treated units.

In order to demonstrate asymptotic properties, two types of asymptotic analysis are carried out: one appropriate when the number of observations at each point in time in each sub-population tends to infinity, and one suitable for stationary aggregate data and in which the number of pre-intervention periods gets large. In this regard, [WHI+15] extends the synthetic control estimator to a cross sectional setting where individual-level data is available and derives its asymptotic distribution when the number of observed individuals goes to infinity. Moreover, [CMM18] propose the *Artificial Counterfactual Estimator (ArCo)*, that is similar in purpose to SCM, and derive its asymptotic distribution when the time dimension is large. However, many of the problems to which the Synthetic Control Method is applied present a cross-section dimension larger than their time dimension, making it impossible to apply the ArCo to them. [Pow16] proposes an inference procedure that uses the gradient of the objective function and relies on the gradient converging to a normally-distributed random variable. This requires asymptotic normality of the estimates for the SCM. Finally, [Li17] derives the asymptotic distribution for the ATE using projection methods, resulting

[1]Basically these papers compute p-values by permuting residuals - for example, [SV18] invert the test statistic to estimate confidence sets for the treatment effect function where the hypothesis testing is carried via a small sample inference procedure for SCM that is similar to Fisher's Exact Hypothesis Test.

in a non-standard asymptotic distribution. However, the analytical asymptotic distribution is hard to obtain and so a sub-sampling method is proposed.

We add to this latter literature, focusing on the case of large number of pre-intervention periods. The work most closely related to ours are [SV18; WHI+15; Li17].

1.1 Abadie, Diamond, and Hainmueller (ADH)

[ADH10] offer no formal inference theory. The idea behind their approach is that a mixture of unaffected units can often provide a better comparison for the treated subject than any single unit could alone.

The framework is based on the Rubin's potential outcomes setup. Let there be T time periods indexed by $t = 1, \ldots, T$ and N sub-populations indexed by $n = 0, 1, \ldots, N$ Let an intervention occur at time period T_0 affecting only group 0, the remaining groups will constitute the control units. Let (y^0_{tn}, y^1_{tn}) be the potential outcomes that would have been observed for unit n at time t without and with exposure to treatment. So that the observed outcome can be written as

$$y_{tn} = D_{tn} y^1_{tn} + (1 - D_{tn}) y^0_{tn}$$

where

$$D_{tn} = \begin{cases} 1 & \text{if } t \geq T_0, n = 0 \\ 0 & \text{otherwise} \end{cases}$$

The difference $\tau_{tn} \equiv y^1_{tn} - y^0_{tn}$ for $t \geq T_0$ will be the treatment effect from intervention for the unit n. The problem comes when estimating the counterfactual y^0_{t0} for $t \geq T_0$.

The key assumption in SCM is the following:

Assumption 1.1. *There exists weights $\beta_n \in [0, 1]$ for $n = 1, \ldots, N$ such that*

$$y^0_{t0} = \sum_{n=1}^{N} \beta_n y^0_{tn}$$

for $t = 1 \ldots, T$ and where the weights sum to one: $\sum_{n=1}^{N} \beta_n = 1$.

Therefore, the ATE (for the treated unit) at $t = T_0 + 1 \ldots, T$ is given by

$$\tau_t = y^1_{t0} - \sum_{n=1}^{N} \beta_n y^0_{t0}$$

and the overall ATE is

$$\tau = \frac{1}{T - T_0 - 1} \sum_{t=T_0+1}^{T} \tau_t$$

The previous assumption says that the outcome of the treated sub-population can be written as a *stable* weighted average of the outcomes of the control sub-populations. The rationale of imposing non-negativity restriction is that in most applications, y_{nt} are positively correlated with each other, and therefore they tend to move up or down together. While the add-to-one restriction implicitly assumes that a weighted average outcomes for the control units and the treated unit's outcome would have followed parallel paths over time in the absence of treatment.

Let $x_t \equiv (y_{t1}, \ldots, y_{tN})^\mathsf{T}$ be a vector of the control unit's outcomes. The most straightforward estimation procedure for β is to solve the minimization problem based on the regression model

$$y_{t0} = \beta^\mathsf{T} x_t + u_{t0} \qquad t = 1, \ldots, T_0 \tag{1.1}$$

i.e.

$$\begin{aligned} &\min \sum_{t=1}^{T_0} (y_{t0} - \beta^\mathsf{T} x_t)^2 \\ s.t. \quad & \\ & ||\beta||_1 = 1 \\ & \beta_i \geq 0 \quad i = 1, \ldots, n \end{aligned} \tag{1.2}$$

We now state the necessary assumptions needed to perform inference analysis for the SCM ATE. Through the rest of this thesis we will work under this framework, though we only need Assumptions (1.1, 1.2, 1.3) for Chapter 4.

Assumption 1.2. *The data* $\{x_t\}_{t=1}^T$ *follows a weakly stationary process*

Assumption 1.3. $\{u_{t0}\}_{t=1}^T$ *is zero-mean and serially uncorrelated satisfying*

$$T_0^{-1/2} \sum_{t=1}^{T_0} x_t u_{t0} \overset{Asy}{\sim} \mathcal{N}(0, \Sigma_1)$$

where $\Sigma_1 = \lim_{T_0 \to \infty} \frac{1}{T_0} \sum_{t=1}^{T_0} \sum_{s=1}^{T_0} \mathbb{E}[u_{t0} u_{s0} x_t x_s^\mathsf{T}]$

Assumption 1.4. *Define* $\nu_{t0} \equiv \tau_t - \tau + u_{t0}$. $\{\nu_{t0}\}_{t=1}^T$ *is zero-mean and satisfies the limit theorem*

$$\frac{1}{\sqrt{T - T_0 - 1}} \sum_{t=T_0+1}^{T} \nu_{t0} \overset{Asy}{\sim} \mathcal{N}(0, \Sigma_2)$$

where $\Sigma_2 = \lim_{(T-T_0-1) \to \infty} \frac{1}{T-T_0-1} \sum_{t=T-T_0-1}^{T} \sum_{s=T-T_0-1}^{T} \mathbb{E}[\nu_{t0} \nu_{s0}]$

Assumption 1.5. *Let $w_t \equiv (y_{t1}, y_{t1}, \ldots, y_{tN}, \tau_t D_{t0})$ for $t = 1, \ldots, T$, where D_{t0} is the post-treatment dummy. $\{w_t\}_{t=1}^{T}$ is a weakly stationary ρ-mixing process with ρ-mixing coefficients $\rho(\tau) = O(\lambda^\tau)$ for some constant $0 < \lambda < 1$.*

These assumptions basically guarantee that the LLN holds, a central limit theorem applies to both the OLS (unrestricted) estimator for (1.1), and to a partial sum of the SCM ATE, and that the estimator $\hat{\beta}$ that solves (1.2) using the pre-treatment data is asymptotically independent with a quantity that involves the post-treatment sample average of the de-mean treatment effects and the idiosyncratic error. While we do not make it explicit, we can further suppose that a consistent estimator for β exists.

Note that knowing the asymptotic behaviour for β will immediately yield the asymptotic behaviour for the ATE τ, thus we focus on the former. In other words, provided we can derive the asymptotic distribution of $\sqrt{T_0}(\hat{\beta}-\beta)$, the asymptotic distribution of $\sqrt{T-T_0-1}(\hat{\tau}-\tau)$ can be found appealing to the following theorem due to [Li17].

Theorem 1.1. *Under assumptions (1.2)-(1.5) we have*

$$\sqrt{T-T_0-1}(\hat{\tau}-\tau) \overset{Asy}{\sim} -\eta \mathbb{E}[x_t^{\mathsf{T}}] Z_1 + \mathcal{N}(0, \Sigma_2)$$

where $\eta \equiv \lim_{T_0,(T-T_0-1)} \sqrt{\frac{T-T_0-1}{T_0}}$, $\sqrt{T_0}(\hat{\beta}-\beta) \overset{Asy}{\sim} Z_1$ is independent with $\mathcal{N}(0, \Sigma_2)$, and

$$\Sigma_2 = \lim_{(T-T_0-1)\to\infty} \frac{1}{T-T_0-1} \sum_{t=T-T_0-1}^{T} \sum_{s=T-T_0-1}^{T} \mathbb{E}[\nu_{t0}\nu_{s0}]$$

On the same line, if we know the confidence region for β, we can derive the confidence region for the ATE τ_t. Hereafter, we thus focus on the asymptotic behaviour of $\sqrt{T_0}(\hat{\beta}-\beta)$, and the confidence region for β.

1.2 Doudchenko, and Imbens

In [DI16], the authors identify a common structure between several methods, which include SCM as a special case. They later on generalize the method, allowing for a permanent additive difference between treated and control units by permitting negative weights and using regularization[2] to deal with potentially large number of possible control units.

Relaxing assumption 1.1, will allow us for a more general linear structure for the imputation of the unobserved:

Assumption 1.6. *There exists weights $\beta_n \in \mathbb{R}$ for $n = 1, \ldots, N$, and a permanent additive difference $\mu \in \mathbb{R}$ such that*

$$y_{t0}^0 = \mu + \sum_{n=1}^{N} \beta_n y_{tn}^0$$

[2]This regularization can be regarded as a robust approach to a least squares estimation.

for $t = 1 \ldots, T$

Adding restriction to μ and β will recover several popular estimators[3].

The estimators $(\hat{\mu}, \hat{\beta})$ solve:

$$\min \sum_{t=1}^{T_0} (y_{t0} - \mu - \beta^{\mathsf{T}} x_t)^2 \tag{1.3}$$

in other words, an ordinary least squares problem. This approach may be attractive in settings where the number of pre-treatment outcomes is large relative to the number of control units, but would be less so in cases where they are of similar magnitude. In any case Doudchenko & Imbens recommend a need for regularization on the weights β. Specifically they recommend to estimate the weights and intercept as an elastic-net least squares regularization:

$$\min ||y_0 - \mu - \beta^{\mathsf{T}} x||_2^2 + \lambda \left(\frac{1-\alpha}{2} ||\beta||_2^2 + \alpha ||\beta||_1 \right) \tag{1.4}$$

and tune the parameters (λ, α) with Cross-Validation.

As in ADH, the way inference is performed is via permutation tests, but no formal inference analysis is carried out[4].

[3]For more information see [DI16]

[4]Recall that there is no rigorous inference theory for the synthetic control estimator if both the numbers of treated and control units are fixed and finite, but both the pre and post-treatment periods are large.

Chapter 2

Projection and tangent cones

2.1 Kathleen Li's projection method

Under the same framework of the previous chapter, [Li17] uses projection methods to derive the asymptotic distribution of the ATE. Unlike all other methodologies, the asymptotic distribution is non-normal and non-standard.

Let Λ be a convex set, consider the minimization problem:

$$\min \sum_{t=1}^{T_0} (y_{t0} - \mu - \beta^\mathsf{T} x_t)^2 \tag{2.1}$$

$$s.t.$$

$$\beta \in \Lambda \tag{2.2}$$

Letting $\Lambda = \Lambda_{SCM} = \{\beta \in \mathbb{R}^N \mid \beta_i \geq 0 \ \& \ ||\beta||_1 = 1\}$ we recover the SCM estimator.

[Li17] modifies the synthetic control method, dropping the add-to-one restriction, but keeping the coefficients non-negative. The rationale is that the treated unit and the control units may exhibit substantial heterogeneity and the treated unit's outcome and a weighted average (with weights sum to one) of the control units' outcomes may not follow parallel paths in the absent of treatment. So that the convex set in (2.1) is:

$$\Lambda_{KL} = \{\beta \in \mathbb{R}^N \mid \beta_i \geq 0\}$$

Let X be the $T_0 \times N$ matrix with its i^{th} row given by x_t^T. Define the projections onto a convex set Λ as

$$\Pi_{\Lambda,T_0}\beta = \text{argmin}_{\lambda\in\Lambda}(\beta - \lambda)^\mathsf{T} \left(XX^\mathsf{T}/T_0\right)(\beta - \lambda)$$

$$\Pi_{\Lambda}\beta = \text{argmin}_{\lambda\in\Lambda}(\beta - \lambda)^\mathsf{T}\mathbb{E}[x_t x_t^\mathsf{T}](\beta - \lambda)$$

The following theorem due to [Li17] establish the asymptotic theory under this setting.

Theorem 2.1. *Let $\hat{\beta}_P$ be a solution to (2.1) and $\hat{\beta}_{OLS}$ a solution to (1.3). Let Z_1 denote the limiting distribution of $\sqrt{T_0}(\hat{\beta}_{OLS} - \beta)$, then*

$$\hat{\beta}_P = \Pi_{\Lambda, T_0} \hat{\beta}_{OLS}$$

and

$$\sqrt{T_0}(\hat{\beta}_P - \beta) \overset{Asy}{\sim} \Pi_{T_{\Lambda,\beta}}$$

where $T_{\Lambda,\beta} := \overline{\cup_{\alpha \geq 0} \alpha\{\Lambda - \Pi_\Lambda \beta\}}$ is the tangent cone of the set Λ at β, and Λ can be either Λ_{SCM} or Λ_{KL}.

Although one can use projection theory to characterize the asymptotic distribution of $\sqrt{T_0}(\hat{\beta}_P - \beta)$, the inference is not straightforward as one has to know β in order to calculate the tangent cone $T_{\Lambda_{KL},\beta}$. A straightforward solution is to plug-in the consistent estimator for β. However, it is an open question if this guarantees desirable properties of the asymptotics. The asymptotic distribution of the synthetic control coefficient estimators depends on whether the true parameters take value at the boundary or not. In practice we do not know which constraints are binding and which are not, making it more difficult to use the asymptotic theory for inference. Moreover, it is known that when parameters fall to the boundary of the parameter space, the standard bootstrap method does not work: [And00]. [Li17] propose a solution to this problem by a sub-sampling method.

Chapter 3

Generalized method of moments (GMM)

Consider M observations $\{Y_i\}_{i=1}^M$ where the Y_i data is generated by a weakly stationary and ergodic process.

The d moment conditions for a vector-valued function $h(X, Y, \cdot) : \mathbb{R}^d \longmapsto \mathbb{R}^d$ is

$$\mathbb{E}[h(X, Y, \theta^*)] = 0$$

and we further consider the null

$$r(\theta^*) = 0$$

The restricted GMM estimator can be written as

$$\begin{aligned}\hat{\theta} = \arg\min_{\theta} & \left(\frac{1}{M}\sum_{i=1}^{M} g(Y_i, \theta)\right)^{\mathsf{T}} \hat{W}_M \left(\frac{1}{M}\sum_{i=1}^{M} g(Y_i, \theta)\right) \\ & s.t. \\ & r(\theta) = 0\end{aligned}$$

where $\hat{W}_M \xrightarrow{p} W$ and W is a positive-definite weighting matrix. Under some regularity conditions[1] the estimator is asymptotically normal[2]

[1] For this we refer to [Hal05].

[2] With the right choice of the weighting matrix W the estimator is also asymptotically efficient. Efficiency is achieved when $W^{-1} = \Omega^{-1} = \mathbb{E}[g(Y, \theta)g(Y, \theta)^{\mathsf{T}}]$, so a good candidate for $\hat{W}_M$ is $\Omega_M(\theta)^{-1}$, where

$$\Omega_M(\theta) = \frac{1}{M}\sum_{i=1}^{M} g(Y_i, \theta)g(Y_i, \theta)^{\mathsf{T}}$$

We subsequently use this as the weighting matrix.

Theorem 3.1. *Under standard regularity assumptions and where* $\Omega_M^{-1} \xrightarrow{p} \Omega^{-1}$*, then*

$$\sqrt{M}(\hat{\theta} - \theta) \overset{Asy}{\sim} \mathcal{N}(0, \Sigma_R)$$

where $\Sigma_R = \Sigma - \Sigma R^\mathsf{T}(R\Sigma R^\mathsf{T})^{-1} R\Sigma$*, and* $\Sigma = (G^\mathsf{T} W G)^{-1}$ *with* G *and* R *the Jacobian matrix*

$$G(\theta) = -\mathbb{E}[\nabla_\theta g(Y, \theta)] \qquad R(\theta) = \nabla_\theta r(\theta)$$

The asymptotic variance-covariance matrix of the restricted GMM estimator is always smaller than or equal to that of the unrestricted GMM estimator, this follows since $\Sigma - \Sigma_R$ is positive semi-definite. This result simply reflects the fact that the restricted GMM estimator, by incorporating more information from the restriction, is more (asymptotically) efficient.

3.1 Inference with GMM

The computation of the weights to construct the synthetic control can be formulated as a GMM problem:

$$\begin{aligned} &\mathbb{E}[(Y - X^\mathsf{T}\beta)X] = 0 \\ &\quad s.t. \\ &\quad r(\beta) = 0 \end{aligned} \tag{3.1}$$

where $r : \mathbb{R}^N \longmapsto \mathbb{R}$ is defined as

$$r(\beta) = \left[1 - ||\beta||_1\right]$$

We apply theorem 3.1 to produce the asymptotic behaviour of the constrained GMM estimator.

Theorem 3.2 (GMM asymptotic variance of SCM)**.** *Under standard regularity assumptions and where* $\Omega_{T_0}^{-1} \xrightarrow{p} \Omega^{-1}$*, then*

$$\sqrt{T_0}(\hat{\beta} - \beta) \overset{Asy}{\sim} \mathcal{N}(0, \Sigma_R)$$

where $\Sigma_R = \Sigma - \Sigma R^\mathsf{T}(R\Sigma R^\mathsf{T})^{-1} R\Sigma$*, with* $\Sigma = \mathbb{E}[x_t x_t^\mathsf{T}]^{-1} \operatorname{Var}[x_t u_t] \mathbb{E}[x_t x_t^\mathsf{T}]^{-1}$ *the heteroskedasticity consistent variance estimator, and*

$$R = \left[-e_{1\times N}^\mathsf{T}\right]$$

with $e = (1, \ldots, 1)^\mathsf{T}$*.*

Note that this GMM estimator might be underidentified. In GMM, identification is essential. Unless parameters are identified, no consistent estimator will exist. In the next chapter, as no uniqueness in solution is to be imposed, we can still perform asymptotic analysis.

Contrasting this inference procedure, [WHI+15] proves that the synthetic control estimator is CAN under the assumption that $(y_{t0}, y_{t1}, \dots, y_{tN})$ is ergodic and stationary.

Theorem 3.3 ([WHI+15]). *Let $\hat{\beta}$ be a solution to (3.1). Suppose that $\{x_t\}$ is ergodic and stationary, and that $\{y_t u_t\}$ satisfies Gordin's conditions*[3]*, then*

$$\sqrt{T_0 - 1}(\hat{\beta} - \beta) \xrightarrow{D} \mathcal{N}(0, C\Gamma C)$$

where

$$C = \mathbb{E}[x_t x_t^\mathsf{T}]^{-1} - \mathbb{E}[x_t x_t^\mathsf{T}]^{-1} e (e^\mathsf{T} \mathbb{E}[x_t x_t^\mathsf{T}]^{-1} e)^{-1} e^\mathsf{T} \mathbb{E}[x_t x_t^\mathsf{T}]^{-1}$$

with $e = (1, \dots, 1)^\mathsf{T}$, and Γ is the long run covariance matrix[4] *of u_t, i.e.*

$$\Gamma = \sum_s \Gamma_s$$

with $\Gamma_s = \mathbb{E}[u_t u_{t-s} x_t x_{t-s}^\mathsf{T}]$.

Stationarity and ergodicity are undoubtedly very strong restrictions. If data is nonstationary, even when imposing a parametric model as in ADH

$$y_{t0,i} = \delta_t + \theta_t Z_i + \lambda_t \mu_i + \epsilon_{ti}$$

where δ_t is an unknown common factor with constant factor loadings, Z_i is a vector of observed covariates not affected by the intervention, θ_t a vector of unknown parameters, λ_t a vector of unobserved factors, μ_i a vector of unknown factor loadings, and ϵ_{ti} unobserved transitory shocks with zero mean - CAN cannot be directly obtained.

[3] See [Hay11].

[4] Methods for estimating Γ are well known, for example the Newey-West estimator.

Chapter 4

Robust Wasserstein Profile Inference

This chapter is based on [BKM19]. Given an estimating equation

$$\mathbb{E}\left[h(W;\theta)\right] = 0$$

we can consider different robust formulations that will yield a set of plausible estimates for the parameter θ - informally this can be thought of as a *confidence region* for $\hat{\theta}$.

Formally, we will present the asymptotic properties of an object called RWP, a novel inference methodology which extends the use of methods inspired by Empirical Likelihood to the setting of optimal transport costs. Our originality lies in applying *distributionally robust optimization* (DRO) theory to study and derive asymptotic distribution of estimators. To our knowledge, DRO has never been used to study the asymptotics of an econometric method.

4.1 Robust Wasserstein Profile Function

Consider the following optimization problem, which may arise in estimation of parameters in econometrics.

$$\min_{\theta:G(\theta)\leq 0} \mathbb{E}_{P_{\text{TRUE}}}\left[H(X,Y,\theta)\right] \tag{4.1}$$

for random elements (X,Y) and a convex function $H(X,Y,\cdot)$ defined over the convex region $\{\theta : G(\theta) \leq 0\}$ and $G : \mathbb{R}^d \mapsto \mathbb{R}$ convex, and where P_{TRUE} denotes the true model. Typically the 'true' measure is approximated by the empirical measure P_n in which case we will denote $\hat{\theta}_n^{ERM}$ to any solution of (4.1) with the empirical measure.

This model may be unknown or too difficult to work with. Therefore, we introduce a proxy P_0 which provides a good trade-off between tractability and model fidelity. So we consider the following robust optimization problem

$$\min_{\theta:G(\theta)\leq 0} \max_{\mathcal{D}_c(P,P_n)\leq\lambda} \mathbb{E}_P\left[H(X,Y,\theta)\right] \tag{4.2}$$

Here P_n is the empirical measure[1], $\mathcal{D}_c$ is defined to be the Wasserstein distance function[2] with cost c, and δ is called the *distributionally uncertainty size*. We will refer as $\hat{\theta}_n^{DRO}$ to any solution of (4.2). Note that $\mathcal{D}_c(P,P_n)\leq\delta$ will define an uncertainty region around the empirical model P_n, we will denote it by $\mathcal{U}_\delta(P_n)=\{P \mid \mathcal{D}_c(P,P_n)\leq\delta\}$. This will ultimately capture the uncertainty in our estimation procedure. For every plausible model $P\in\mathcal{U}_\delta(P_n)$ there is an optimal choice of parameter θ^* such that minimizes $\mathbb{E}_P\left[H(X,Y,\theta)\right]$. The set of all such parameters will be denoted by

$$\Delta_n(\delta) := \{\theta(P) \,:\, \theta\in \operatorname{argmin}_\theta \mathbb{E}_P[H(X,Y,\theta)] \;\; P\in\mathcal{U}_\delta(P_n)\}$$

The problem now translates to finding δ such that

$$\theta^*\in\Delta_n(\delta)$$

with probability at least $(1-\alpha)$, where α is set to be the confidence level.

Suppose that solutions to (4.1) are given by a system of equations of the form

$$\mathbb{E}_{P_n}\left[h(X,Y,\theta)\right]=0$$

for a suitable $h(\cdot)$.

The Robust Wasserstein Profile (RWP) function as defined by [BKM19] is then

$$R_n(\theta) := \inf\{\mathcal{D}_c(P,P_n) \;:\; \mathbb{E}_P\left[h(X,Y,\theta)\right]=0\} \tag{4.3}$$

The following proposition is a key observation which will lead to the construction of confidence region in parameter estimation.

Proposition 4.1. *Let $\chi_{1-\alpha}$ be the $(1-\alpha)$ quantile of the function $R_n(\theta)$. Then $\Delta_n(\chi_{1-\alpha})$ is a $(1-\alpha)$ confidence region for θ.*

Proof. The $1-\alpha$ quantile for the RWP function is given by:

$$\chi_{1-\alpha}=\inf\{z \mid P(R_n(\theta)\leq z)\geq 1-\alpha\}$$

[1] and whose weak limit is P_{TRUE}.

[2] Let the cost function satisfy $c(x,y)\longmapsto[0,\infty)$. Define

$$\mathcal{D}_c(\mu,\nu) := \inf_{\gamma\in\Gamma(\mu,\nu)} \int c(x,y)d\gamma(x,y)$$

where $\Gamma(\mu,\nu)$ denotes the collection of all measures with marginal μ and ν on the first and second factors respectively.

The definition of the RWP function allows us to write $\Delta_n(\chi_{1-\alpha})$ as

$$\Delta_n(\chi_{1-\alpha}) = \{\theta \mid R_n(\theta) \leq \chi_{1-\alpha}\}$$

Therefore,

$$P(\theta \in \Delta_n(\chi_{1-\alpha})) = P(R_n(\theta) \leq \chi_{1-\alpha}) = 1 - \alpha$$

so $\Delta_n(\chi_{1-\alpha})$ is a $(1-\alpha)$ confidence region for θ. □

Remark 4.1. *Proposition 8 of [BKM19] establishes a min-max theorem for the DRO formulation:*

$$\min_{\theta:G(\theta)\leq 0} \max_{\mathcal{D}_c(P,P_n)\leq\lambda} \mathbb{E}_P\left[H(X,Y,\theta)\right] = \max_{\mathcal{D}_c(P,P_n)\leq\lambda} \min_{\theta:G(\theta)\leq 0} \mathbb{E}_P\left[H(X,Y,\theta)\right]$$

This indicates that $\hat{\theta}_n^{DRO} \in \Delta_n(\delta)$, otherwise the left hand side of the equation above would be strictly larger than the right hand side. Trivially, $\hat{\theta}_n^{ERM}$ is also inside $\Delta_n(\delta)$. This property is an attractive feature, as this confidence region will also include 'Doudchenko-Imbens' estimators[3], i.e. those who solve (1.4).

The following proposition due to [BKM19] gives a dual formulation for the RWP function, which is useful to derive its asymptotic properties, and easier to compute as the problem passes to have an infinite dimensional formulation to a finite dimensional one.

Theorem 4.2 ([BKM19])**.** *Let $h(\cdot,\theta)$ be Borel measurable, and $\Omega = \{(u,w) \in \mathbb{R}^m \times \mathbb{R}^m : c(u,w) < \infty\}$ be Borel measurable and non empty. Further, suppose that 0 lies in the interior of the convex hull of $\{h(u,\theta) : u \in \mathbb{R}^m\}$. Then,*

$$R_n(\theta) = \sup_{\lambda\in\mathbb{R}^r} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \sup_{u\in\mathbb{R}^m} \{\lambda^\mathsf{T} h(u,\theta) - c(u,W_i)\} \right\}$$

Remark 4.2. *Note that it might be computationally costly or unfeasible to derive the $1-\alpha$ quantile of the function $R_n(\theta^*)$, so instead we will focus on its asymptotic distribution.*

The following theorem gives the asymptotic distribution of the RWP function

Theorem 4.3 ([BKM19])**.** *Consider the cost function[4] associated with the Wasserstein distance (and hence with the RWP function), to be*

$$c\left((x,y),(u,v)\right) = \begin{cases} ||x-u||_2 & \text{if } y = v \\ \infty & \text{otherwise} \end{cases}$$

Suppose that

[3] It can be proven with the techniques presented in [BKM19], that problem (1.4) can be formulated as a DRO problem.

[4] As this modified cost function assigns infinite cost when $y \neq v$, the infimum of the RWP function is effectively over joint distributions that do not alter the marginal distribution of Y. As a consequence, the resulting uncertainty set $\mathcal{U}_\delta(P_n)$ admits distributional ambiguities only with respect to the predictor variables X.

(i) $\theta^* \in \mathbb{R}^d$ *satisfies* $\mathbb{E}\left[h((X,Y),\theta^*)\right] = 0$ *and* $\mathbb{E}||h((X,Y),\theta^*)||_2^2 < \infty$

(ii) For each $\zeta \neq 0$*, the partial derivative* $D_x h((x,y),\theta^*)$ *exists, is continuous, and satisfies,*

$$P(||\zeta^\mathsf{T} D_x h((X,Y),\theta^*)||_2 > 0) > 0$$

(iii) Assume that there exists $\bar{\kappa} : \mathbb{R}^m \longmapsto [0,\infty)$ *such that*

$$||D_x h(x+\Delta, y, \theta^*) - D_x h(x,y,\theta^*)||_2 \leq \bar{\kappa}(x,y)||\Delta||_2$$

for all $\Delta \in \mathbb{R}^d$*, and* $\mathbb{E}[\bar{\kappa}(X,Y)^2] < \infty$*.*

Then,

$$nR_n(\theta^*) \overset{Asy}{\sim} \bar{R}(2)$$

where

$$\bar{R}(2) := \sup_{\zeta \in \mathbb{R}^d} \left\{ 2\zeta^\mathsf{T} H - \mathbb{E}||\zeta^\mathsf{T} D_x h((X,Y),\theta^*)||_2^2 \right\}$$

with $H \sim \mathcal{N}(0, \mathrm{cov}[h((X,Y),\theta^*)])$

For further details in the RWP function, its properties and connection with estimating literature, we refer to [BKM19], and [BK17], and the references therein.

It is important to mention that the attempt in this thesis is to derive the exact uncertainty set Δ. In [BKS19] a theorem is presented giving the asymptotic normality of underlying DRO estimators.

Theorem 4.4. *Suppose that*

(i) $H(\cdot)$ *is twice continuously differentiable, non-negative, and for each* (X,Y)*,* $H(X,Y,\cdot)$ *is convex.*

(ii) $\theta^* \in \mathbb{R}^d$ *satisfies* $\mathbb{E}\left[h((X,Y),\theta^*)\right] = 0$ *and* $\mathbb{E}||h((X,Y),\theta^*)||_2^2 < \infty$

(iii) Both $\mathbb{E}[D_\theta h(X,Y,\theta^*)]$*, and* $\mathbb{E}[D_x h(X,Y,\theta^*) D_x h(X,Y,\theta^*)^\mathsf{T}]$ *are strictly positive definite.*

(iv) Assume that there exists $\kappa, \kappa', \bar{\kappa} : \mathbb{R}^m \longmapsto [0,\infty)$ *such that*

$$||D_x h(x+\Delta, y, \theta^*) - D_x h(x,y,\theta^*)||_2 \leq \kappa(x,y)||\Delta||_2$$

$$||D_x h(x+\Delta, y, \theta^*+u) - D_x h(x,y,\theta^*)||_2 \leq \bar{\kappa}(x,y)\left(||\Delta||_2 + ||u||_2\right)$$

$$||D_x h(x+\Delta, y, \theta^*+u) - D_\theta h(x,y,\theta^*)||_2 \leq \kappa'(x,y)\left(||\Delta||_2 + ||u||_2\right)$$

and $\mathbb{E}[\kappa(X,Y)^2] < \infty$*,* $\mathbb{E}[\bar{\kappa}(X,Y)^2] < \infty$*,* $\mathbb{E}[\kappa'(X,Y)^2] < \infty$*.*

Let $\mathbb{E}||(X,Y)||^2 < \infty$, $C := \mathbb{E}[D_\theta h(X,Y,\theta^*]$, *and* $H \sim \mathcal{N}(0, \text{cov}[h((X,Y),\theta^*)])$; *then*

$$\left(\sqrt{n}(\theta_n^{ERM} - \theta^*),\ \sqrt{n}(\Delta_n(n^{-1}\delta) - \theta^*)\right) \overset{Asy}{\sim} \left(C^{-1}H,\ \Phi(\delta) + C^{-1}H\right)$$

where $\Phi(\delta) := \{z \mid \sup_\zeta\{\zeta^\mathsf{T} Cz - \frac{1}{4}\mathbb{E}||\zeta^\mathsf{T} D_x h(X,Y,\theta^*)||^2\} \leq \delta\}$.

Note that, by the continuous mapping theorem, we have the approximation:

$$\Delta_n(n^{-1}\delta) \approx \hat{\theta}_n^{ERM} + n^{-1/2}\Phi(\delta)$$

the following result is the basis of constructing the asymptotic confidence region.

Theorem 4.5. *Let* C_n *be a consistent estimator of* $C = \mathbb{E}[D_\theta h\,(X,Y,\theta^*)]$, *and* $\delta(n) = \delta + o(1)$; *then*

$$\Phi_n(\delta(n)) := \{z \mid \sup_\zeta\{\zeta^\mathsf{T} C_n z - \frac{1}{4}\mathbb{E}_{P_n}||\zeta^\mathsf{T} D_x h(X,Y,\theta^*)||^2\} \leq \delta\} \Longrightarrow \Phi(\delta)$$

We will use this results in the next section to derive the asymptotics of the SCM estimator.

4.2 Inference via the RWP function

Proposition (4.1), and an application of Theorem (4.3) will be the basis to derive an exact confidence region for β. On the other hand, Theorems (4.4), and (4.5), will give the asymptotic behaviour of this confidence region.

The empirical risk minimization problem that compute the synthetic control weights is:

$$\min_{\beta\,:\,||\beta||_1 = 1\ \beta_i \geq 0} \mathbb{E}_{P_{T_0}}||Y - X^\mathsf{T}\beta||^2 \tag{4.4}$$

Note that the KKT conditions are

$$\begin{aligned}(y - \beta^\mathsf{T} x)x - \lambda e + \mu &= 0\\ 1 - ||\beta||_1 &= 0\\ \beta - s^2 &= 0\\ \text{diag}(\mu)\,\text{diag}(s) &= 0\end{aligned}$$

where $\lambda \in \mathbb{R}$, $e = (1,\ldots,1)^\mathsf{T} \in \mathbb{R}^N$, $\mu = (\mu_1,\ldots,\mu_N)^\mathsf{T}$, and $s = (s_1,\ldots,s_N)$.

We define $h(x,y;\beta,\lambda,\mu,s) : \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^{N+1+N+N} \longmapsto \mathbb{R}^{N+1+N+N}$ to be

$$h(x, y, \beta, \lambda, \mu, s) = \begin{bmatrix} (y - \beta^{\mathsf{T}} x)x - \lambda e + \mu \\ 1 - ||\beta||_1 \\ \beta - s^2 \\ \operatorname{diag}(\mu) \operatorname{diag}(s) \end{bmatrix} \tag{4.5}$$

and apply Theorem 4.3 to this function, to derive our main result.

Theorem 4.6. *Consider $h(x, y, \beta, \lambda, \mu, s)$ as defined by (4.5) For $\beta \in \mathbb{R}^N$ let*

$$R_{T_0}(\beta) = \inf\{\mathcal{D}_c(P, P_{T_0}) \ : \ \mathbb{E}_P[h(X, Y, \beta, \lambda, \mu, s] = 0\}$$

where the cost function is

$$c\left((x, y), (u, v)\right) = \begin{cases} ||x - u||_2 & \text{if } y = v \\ \infty & \text{otherwise} \end{cases}$$

Under the null hypothesis that the training samples $\{(X_i, Y_i)\}_i$ are obtained independently from a constrained model $Y = \beta^{*\mathsf{T}} X + u$ where $||\beta^*||_1 = 1$, and $\beta_i^* \geq 0$. The error term u has zero mean and variance σ^2, and $\Sigma = \mathbb{E}[XX^{\mathsf{T}}]$ is invertible. Then,

$$T_0 R_{T_0}(\beta^*) \overset{Asy}{\sim} \bar{R}$$

where

$$\bar{R} = \mathcal{N}(0, A)^{\mathsf{T}} \left[\sigma^2 Id - (\lambda^* e - \mu^*)\beta^{*\mathsf{T}} - \beta^*(\lambda^* e - \mu^*)^{\mathsf{T}}\right]^{-1} \mathcal{N}(0, A)$$

and $A = \sigma^2\Sigma - \lambda^{*2} ee^{\mathsf{T}} + \lambda^* e\mu^{*\mathsf{T}} + \lambda^*\mu^* e^{\mathsf{T}} - \mu^*\mu^{*\mathsf{T}}$

Proof. To show that the RWP function converges in distribution, we verify the assumptions of Theorem (4.3) with $h(\cdot)$ defined in (4.5).

Under the null hypothesis, the KKT conditions are satisfied together with the slackness complementarity conditions, therefore

$$\mathbb{E}[h(X, Y; \beta^*)] = \begin{bmatrix} uX - \lambda^* e + \mu^* \\ 1 - ||\beta^*||_1 \\ \beta^* - s^{*2} \\ \operatorname{diag}(\mu^*) \operatorname{diag}(s^*) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

and by the triangle inequality

$$\mathbb{E}||h(X, Y; \beta^*)||^2 \leq \mathbb{E}[||uX||^2 + ||\lambda^* e - \mu^*||] = \sigma^2 \mathbb{E}||X||^2 + ||\lambda^* e - \mu^*|| < \infty$$

which is finite because the trace of the matrix Σ is finite. This verifies assumption (i).

Now,

$$D_x h(x, y, \beta^*) = \left[\frac{uId - x\beta^{*\mathsf{T}}}{0} \right]_{(2N+1)\times N}$$

which is clearly continuous and for any $0 \neq (\zeta, \eta) \in \mathbb{R}^N \times \mathbb{R}^{2N+1}$

$$P(||(\zeta, \eta)^\mathsf{T} D_x h(X, Y, \beta^*)||^2 = 0) = P(u\zeta = \zeta^\mathsf{T} X\beta) = 0$$

and thus satisfying assumption (ii). In addition,

$$||D_x h(x + \Delta, y, \beta^*) - D_x h(x, y, \beta^*)|| = ||\beta^{*\mathsf{T}} \Delta Id - \Delta\beta^{*\mathsf{T}}|| \leq c||\Delta||$$

for some positive constant c.

As a consequence of Theorem (4.3)

$$T_0 R_{T_0}(\beta^*) \overset{Asy}{\sim} \sup_{(\zeta,\eta)\in\mathbb{R}^N\times\mathbb{R}^{2n+1}} \left\{ 2(\zeta, \eta)^\mathsf{T} H - \mathbb{E}||(\zeta, \eta)^\mathsf{T} \left[\frac{uId - X\beta^{*\mathsf{T}}}{0} \right]_{(2N+1)\times N} ||^2 \right\}$$

Note that $H \sim \mathcal{N}(0, \operatorname{cov} h(X, Y; \beta^*))$ where

$$\operatorname{cov} h(X, Y; \beta^*) = \mathbb{E}[hh^\mathsf{T}] = \left[\begin{array}{c|c} A_{N\times N} & 0 \\ \hline 0 & 0 \end{array} \right]_{(2N+1)\times(2N+1)}$$

and

$$\begin{aligned} A &= \mathbb{E}[u^2 XX^\mathsf{T} + \lambda^* ueX^\mathsf{T} - u\mu^* X^\mathsf{T} + u\lambda^* Xe^\mathsf{T} + \lambda^{*2} ee^\mathsf{T} - \lambda^*\mu^* e^\mathsf{T} - uX\mu^{*\mathsf{T}} - \lambda^* e\mu^{*\mathsf{T}} + \mu^*\mu^{*\mathsf{T}}] \\ &= \sigma^2\Sigma - \lambda^{*2} ee^\mathsf{T} + \lambda^* e\mu^{*\mathsf{T}} + \lambda^*\mu^* e^\mathsf{T} - \mu^*\mu^{*\mathsf{T}} \end{aligned}$$

further note that $-\lambda^{*2} ee^\mathsf{T} + \lambda^* e\mu^{*\mathsf{T}} + \lambda * \mu^* e^\mathsf{T} - \mu^*\mu^{*\mathsf{T}}$ is negative semi-definite.

We will 'partition' the distribution H as

$$H = \left[\begin{array}{c|c} \mathcal{Z} & \delta_{2N+1} \end{array} \right]$$

where $\delta_{2N+1} = (\delta, \ldots, \delta)$ denotes a $2N + 1$-dimensional delta-distribution and $\mathcal{Z} \sim \mathcal{N}(0, A)$.

Therefore the limiting distribution can be simplified to

$$\begin{aligned} T_0 R_{T_0}(\beta^*) &\overset{Asy}{\sim} \sup_{\zeta\in\mathbb{R}^N} \left\{ 2\zeta^\mathsf{T}\mathcal{Z} - \mathbb{E}||\zeta^T(uId - X\beta^{*\mathsf{T}})||^2 \right\} \\ &= \mathcal{Z}^\mathsf{T} \mathbb{E}[(uId - X\beta^{*\mathsf{T}})(uId - X\beta^{*\mathsf{T}})^\mathsf{T}]^{-1} \mathcal{Z} \\ &= \mathcal{Z}^\mathsf{T} \left[\sigma^2 Id - (\lambda^* e - \mu^*)\beta^{*\mathsf{T}} - \beta^*(\lambda^* e - \mu^*)^\mathsf{T} + ||\beta^*||^2\Sigma \right]^{-1} \mathcal{Z} \end{aligned}$$

and it can be easily justified that $\left[\sigma^2 Id - (\lambda^* e - \mu^*)\beta^{*\mathsf{T}} - \beta^*(\lambda^* e - \mu^*)^\mathsf{T} + ||\beta^*||^2\Sigma \right]$ is positive definite and so invertible. □

Remark 4.3. *(i) Note that as in restricted GMM, the variance of the RWP function is less than in the unrestricted setting.*

(ii) Observe that the limiting distribution is a generalized chi-squared distribution[5]*:*

Let

$$B = \sigma^2 Id - (\lambda^* e - \mu^*)\beta^{*\mathsf{T}} - \beta^*(\lambda^* e - \mu^*)^{\mathsf{T}} + ||\beta^*||^2\Sigma$$

and using the spectral theorem let

$$A^{\frac{1}{2}} B^{-1} A^{\frac{1}{2}} = U\Lambda U^{\mathsf{T}}$$

be the eigendescomposition , we have that $N = U^{\mathsf{T}} A^{-\frac{1}{2}} Z$ has standard normal distribution. As a result

$$\bar{R} = \mathcal{Z}^{\mathsf{T}} B^{-1} \mathcal{Z} = N^{\mathsf{T}} \Lambda N = \sum_{i=1}^{N} \lambda_i N_i^2$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$.

To stress the relation of β^* in the RWP asymptotic distribution, we will denote it as $\bar{R}(\beta^*) = N^{\mathsf{T}} \Lambda(\beta^*) N$.

(iii) As [BKM19] conjectures in a LASSO setting, one could aim to achieve lower bias in estimation by working with the $(1-\alpha)$-quantile of the limit law $\bar{R}(\beta^*)$, instead of that of an stochastic upper bound independent of the estimator β^*. In order to do so, they propose to use any consistent estimator for β^* to be plugged in the expression for $\bar{R}$. However, it is an open problem if this plug-in approach indeed enjoys better generalization guarantees - just as in the remark of Theorem (2.1).

Recall from proposition (4.1) that a $(1-\alpha)$ confidence region for the parameter β is given by

$$\Delta_{T_0}(\chi_{1-\alpha}) = \{\beta \mid R_{T_0}(\beta) \leq T_0^{-1} \chi_{1-\alpha}\} \tag{4.6}$$

where $\chi_{1-\alpha}$ is the $(1-\alpha)$-quantile of $\bar{R}(\hat{\beta}) = N^{\mathsf{T}} \Lambda(\hat{\beta}) N$, and $R_{T_0}(\beta)$ can be computed as in Theorem (4.2):

[5] There has been some work on computing things with this distribution: [IMH61] and [Dav80] numerically invert the characteristic function. [SO77] write the distribution as an infinite sum of central chi-squared variables. [LTZ09] approximate it with a noncentral chi-squared distribution based on cumulant matching.

$$R_{T0}(\beta) = \sup_{\lambda \in \mathbb{R}^N} \left\{ -\frac{1}{n} \sum_{i=1}^{N} \sup_{x \in \mathbb{R}^N} \left\{ \lambda^{\mathsf{T}} (Y_i - \beta^{\mathsf{T}}) x - ||x - X_i||^2 \right\} \right\}$$
$$= \sup_{\{\lambda \mid P \text{ is pos def }\}} \left\{ -\frac{1}{n} \sum_{i=1}^{N} \sup_{\{x \mid Px = Y_i + 2X_i\}} \left\{ \lambda^{\mathsf{T}} (Y_i - \beta^{\mathsf{T}}) x - ||x - X_i||^2 \right\} \right\}$$

with $P = 2Id + \lambda\beta^{\mathsf{T}} + \beta\lambda^{\mathsf{T}}$.

In order to solve the inequality (4.6) we propose the following procedure[6].

Consider a second order model for the RWP function $R_{T_0}(\beta)$ around a consistent estimator $\hat{\beta}$. This is a fair approximation since the RWP function is convex and has a global minimum at $\hat{\beta}$:

$$\begin{aligned} R_{T_0}(\beta) &= R_{T_0}(\hat{\beta}) + \nabla R_{T_0}^{\mathsf{T}}(\hat{\beta})(\beta - \hat{\beta}) + \frac{1}{2}(\beta - \hat{\beta})^{\mathsf{T}} \nabla^2 R_{T_0}(\hat{\beta})(\beta - \hat{\beta}) + \mathcal{O}||\beta - \hat{\beta}||^3 \\ &= \frac{1}{2}(\beta - \hat{\beta})^{\mathsf{T}} \nabla^2 R_{T_0}(\hat{\beta})(\beta - \hat{\beta}) + \mathcal{O}||\beta - \hat{\beta}||^3 \\ &\approx \frac{1}{2}(\beta - \hat{\beta})^{\mathsf{T}} \nabla^2 R_{T_0}(\hat{\beta})(\beta - \hat{\beta}) \end{aligned}$$

Therefore $\Delta_{T_0}(\chi_{1-\alpha})$ can be approximated as

$$\begin{aligned} \Delta_{T_0}(\chi_{1-\alpha}) &= \{\beta \mid R_{T_0}(\beta) \leq T_0^{-1}\chi_{1-\alpha}\} \\ &\approx \{\beta \mid \frac{1}{2}(\beta - \hat{\beta})^{\mathsf{T}} \nabla^2 R_{T_0}(\hat{\beta})(\beta - \hat{\beta}) \leq T_0^{-1}\chi_{1-\alpha}\} \end{aligned}$$

This defines an ellipsoid centered at $\hat{\beta}$ where the principal axis are determined by the eigenvectors of the Hessian of the RWP function, and the eigenvalues are the reciprocal of the squares of the semi-axes.

$$\Delta_{T_0}(\chi_{1-\alpha}) \approx \{\beta \mid (\beta - \hat{\beta})^{\mathsf{T}} \nabla^2 R_{T_0}(\hat{\beta})(\beta - \hat{\beta}) \leq 2T_0^{-1}\chi_{1-\alpha}\} \subseteq B(\hat{\beta}, \mathsf{r})$$

where $B(\hat{\beta}, \mathsf{r})$ denotes the ball centered at $\hat{\beta}$ with radius

$$\mathsf{r} = \sqrt{\frac{2\chi_{1-\alpha}}{T_0 \lambda_{\min}(\nabla^2 R)}}$$

[6]Alternative one can consider to solve the inequality approximately, exploiting the convex structure of the RWP function.

where $\lambda_{\min}(\nabla^2 R)$ denotes the minimum eigenvalue of the Hessian $\nabla^2 R_{T_0}(\hat{\beta})$. Finally, for the whole sample period, the outcome y_{t0} is generated by

$$y_{t0} = \beta^{\mathsf{T}} x_t + D_{t0}\tau_t + u_{t0} \qquad t = 1, \ldots, T_0, \ldots T$$

where D_{t0} is the post-treatment dummy, and u_{t0} has variance σ^2 so that

$$\begin{aligned}
|\hat{\tau}_t - \tau_t| =& |y_{t0} - \hat{y}^0_{t0} - \tau_t| = |\beta^{\mathsf{T}} x_t + \tau_t + u_{t0} - \hat{\beta}^{\mathsf{T}} x_t - \tau_t| \\
=& |(\beta - \hat{\beta})^{\mathsf{T}} x_t + u_{t0}| \leq |(\beta - \hat{\beta})^{\mathsf{T}} x_t| + |u_{t0}| \\
\leq& ||\beta - \hat{\beta}||\, ||x_t|| + |u_{t0}| \\
\leq& \mathsf{r}||x_t|| + \sigma z_{(1-\alpha/2)}
\end{aligned}$$

thus, the confidence region for the ATE of the SCM estimator is given by

$$B\left(\hat{\tau}_t, \mathsf{r}||x_t|| + \sigma z_{(1-\alpha/2)}\right)$$

while the confidence region for the overall ATE is given by

$$B\left(\hat{\tau}, \mathsf{r}\left|\left|\frac{1}{T - T_0 - 1}\sum_{t=T_0+1}^{T} x_t\right|\right|\right)$$

Remark 4.4. *Note that this confidence regions can be interpreted à la Manski, meaning that they were derived considering the worst-case scenario.*

We can contrast this procedure with the asymptotic behaviour of the *ERM* estimator and the confidence region given by Theorems 4.4 and 4.5.

Theorem 4.7. *Consider the same framework of Theorem (4.6), and the definitions therein. Then,*

$$\sqrt{T_0}(\hat{\beta} - \beta^*) \overset{Asy}{\sim} \Sigma^{-1}\mathcal{N}(0, A)$$

and

$$\Delta_{T_0}(T_0^{-1/2}\chi_{1-\alpha}) \approx \hat{\beta} + T_0^{-1}\{z \mid z^{\mathsf{T}}\Sigma B^{-1}\Sigma z \leq \chi_{1-\alpha}\}$$

where $\chi_{1-\alpha}$ is the $1-\alpha$ quantile of the RWP asymptotic function.

Lets observe that $\Sigma B^{-1}\Sigma$, as a positive quadratic form, represents an ellipsoid. As before, considering the confidence region à la Manski, we can further approximate the confidence region for β^* as

$$B\left(\hat{\beta}, \sqrt{\frac{\chi_{1-\alpha}}{T_0\lambda_{\min}(\Sigma B^{-1}\Sigma)}}\right)$$

where $\lambda_{\min}(\Sigma B^{-1}\Sigma)$ is the minimum eigenvalue of $\Sigma B^{-1}\Sigma$

and so the confidence region for the ATE of the SCM is given by

$$B\left(\hat{\tau}_t, \sqrt{\frac{\chi_{1-\alpha}}{T_0\lambda_{\min}(\Sigma B^{-1}\Sigma)}}||x_t|| + \sigma z_{(1-\alpha/2)}\right)$$

while the confidence region for the overall ATE is given by

$$B\left(\hat{\tau}, \sqrt{\frac{\chi_{1-\alpha}}{T_0\lambda_{\min}(\Sigma B^{-1}\Sigma)}}\left|\left|\frac{1}{T-T_0-1}\sum_{t=T_0+1}^{T} x_t\right|\right|\right)$$

. □

Chapter 5

Empirical example

In this chapter we illustrate and contrast the methods outlined here with an empirical application. We revisit the classical paper [ADH10], which estimates the effect of Proposition 99, a large-scale tobacco control program that California implemented in 1988. It uses smoking per capita as the outcome and uses a single treated unit (California) and $N = 29$ states without such anti-smoking measures as the set of potential controls. In order to conduct inference, the authors run placebo studies by applying the synthetic control method to states that did not implement a large-scale tobacco control program during the sample period of study. They argue that as the estimated gap between California and its synthetic control is "unusually large relative to the distribution of the gaps for the states in the donor pool", compared to placebo states and their respective synthetic control, the treatment effect is not driven entirely by 'chance' and so they conclude significance. The next figure can be found on [ADH10].

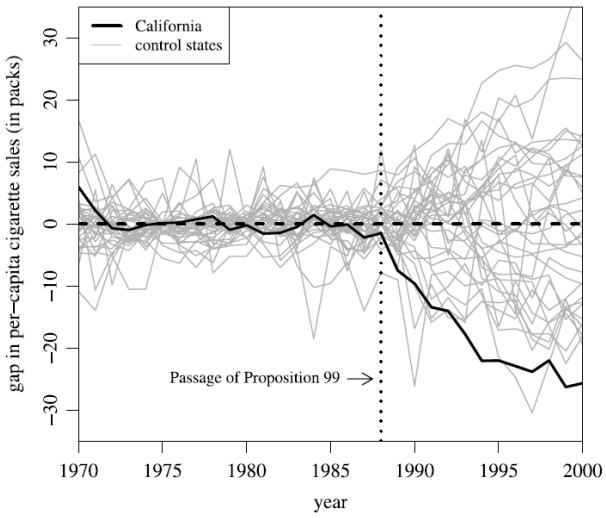
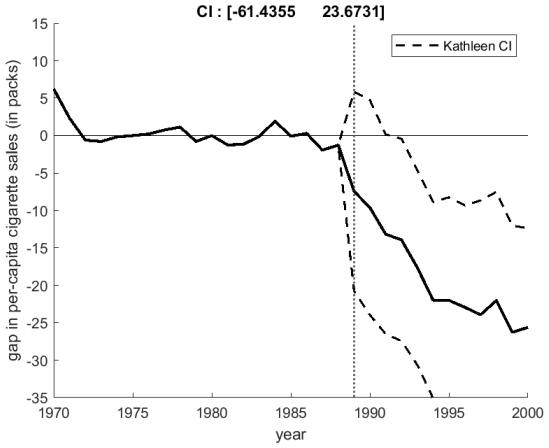
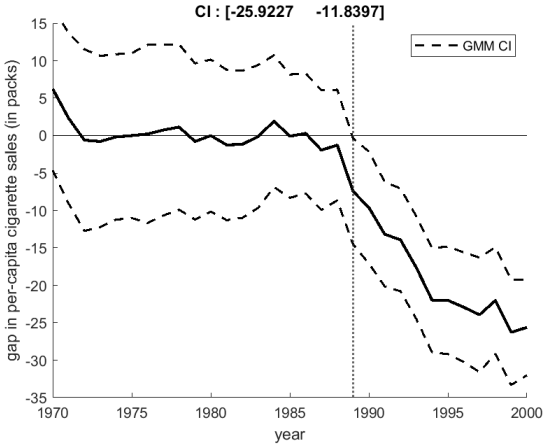


Figure 5.1: Treatment effect with placebos

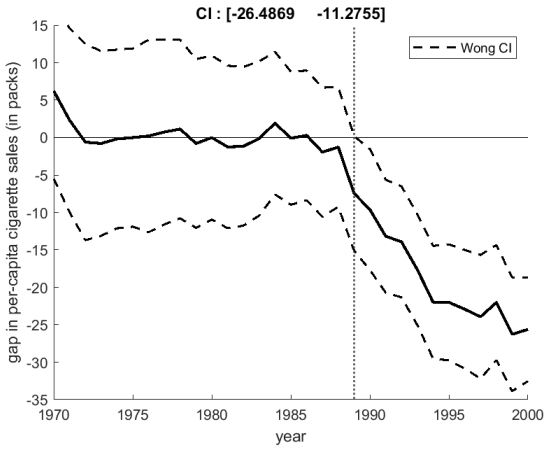
(a) Kathleen sub-sampling



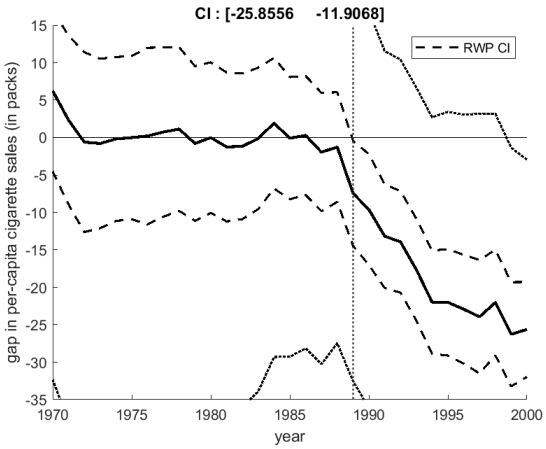
(b) GMM



(c) Wong GMM



(d) RWP CI



(e) Asymptotic RWP CI

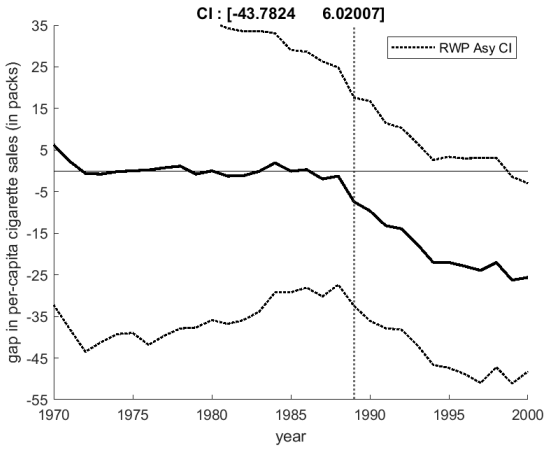


Figure 5.2: Analytical confidence intervals

The figure above contrasts the different inference procedures discussed here. We note that all methods conclude overall significance, excepting Kathleen's sub-sampling procedure (Panel a), and the asymptotic Wasserstein CI (Panel e). We first turn our attention to the GMM-derived confidence intervals: It looks that the GMM CI (Panel b) - Theorem (3.2) - is the most efficient one, and as expected the CI derived with Theorem (3.3) (Panel c) has larger radius as it is a 'robustification' of Theorem (3.2). The CI derived with the asymptotics obtained in Theorem (4.7) (Panel d) are slightly larger but similar to the one obtained with restricted GMM, this is because we can regard the asymptotic distribution of the ERM estimator as a constrained GMM estimator but with a different weighting matrix - which is not optimum in terms of efficiency. Finally, (Panel d) shows the confidence interval using the Wasserstein profile function; it shows both the 'exact' confidence interval and its asymptotic approximation. (Panel e) only displays the asymptotic Wasserstein CI.

Chapter 6

Conclusion

The main contribution of this thesis, besides the analysis of the asymptotics of the SCM estimator - which is in its own a great contribution, is the introduction of the Robust Wasserstein Profile Inference methodology. This is a novel approach, that recovers confidence region for the estimate of interest, thus introducing a new form of making inference in statistics and econometrics. We highlight that this approach may not be easily scalable, as it depends on the form of the RWP function. However, a tractable approach is obtained by considering the asymptotic Wasserstein confidence interval.

Note that the methods outlined in this work can be straightforwardly extended to include covariates to improve the estimation and inference procedure, as well as incorporate heteroskedastic-consistent standard errors. The confidence regions are interpreted à la Manski, meaning they were derived considering the worst-case scenario. This means that the length of the confidence regions can be further tightened.

It is also worth noting that assumption (1.6) can be easily relaxed to allow for any other functional relation, allowing for non-linearities or even non-parametric forms. The RWP methodology is easily adapted to allow for such changes.

As in [ADH10], the computation of the weights can be simplified by considering only a few linear combination of pre-intervention outcomes and checking whether assumption (1.2) holds approximately for the resulting weights. Another possibility, is to modify the 2-norm in the loss function replacing it with a norm induced by a matrix V. The choice of V can be data-driven. One possibility is to choose V among positive definite and diagonal matrices such that the mean squared prediction error of the outcome variable is minimized for the pre-intervention periods (see [AG03], appendix B for details).

We want to emphasize that this methodology to recover confidence interval in SCM can be adapted to any of the extensions mentioned in Chapter 1.

To illustrate the results of the thesis, we revisit the classic paper by [ADH10] and derived

confidence intervals for the treatment effect by each of the methods outlined here. We see that all methods are consistent between them, but some prove to be more efficient than others.

An advantage of the confidence region obtained with the RWP function is that it contains both an empirical risk minimizer and a distributionally robust minimizer - this is an attractive feature as some SCM estimator variant, such as the proposed by [DI16] fall in this confidence region. It is an interesting question to ask which other estimator variants also fall in this confidence region.

Bibliography

[ADH10] Alberto Abadie, Alexis Diamond, and Jens Hainmueller, *Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program*, Journal of the American Statistical Association **105** (2010), no. 490, 493–505.

[ADH15] Alberto Abadie, Alexis Diamond, and Jens Hainmueller, *Comparative politics and the synthetic control method*, American Journal of Political Science **59** (2015), no. 2, 495–510.

[AG03] Alberto Abadie and Javier Gardeazabal, *The economic costs of conflict: A case study of the basque country*, American Economic Review **93** (2003), no. 1, 113–132.

[AI17] Susan Athey and Guido W. Imbens, *The state of applied econometrics: Causality and policy evaluation*, Journal of Economic Perspectives **31** (2017), no. 2, 3–32.

[AJK+16] Daron Acemoglu, Simon Johnson, Amir Kermani, James Kwak, and Todd Mitton, *The value of connections in turbulent times: Evidence from the united states*, Journal of Financial Economics **121** (2016), no. 2, 368 – 391.

[AL18] Alberto Abadie and Jérémy L'Hour, *A penalized synthetic control estimator for disaggregated data.*, Tech. report, 2018.

[And00] Donald W. K. Andrews, *Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space*, Econometrica **68** (2000), no. 2, 399–405.

[ASS18] Muhammad Amjad, Devavrat Shah, and Dennis Shen, *Robust synthetic control*, Journal of Machine Learning Research **19** (2018), no. 22, 1–51.

[BCL+18] Janet Bouttell, Peter Craig, James Lewsey, Mark Robinson, and Frank Popham, *Synthetic control methodology as a tool for evaluating population-level health interventions*, Journal of Epidemiology & Community Health **72** (2018), no. 8, 673–678.

[BK17] Jose Blanchet and Yang Kang, *Distributionally robust groupwise regularization estimator*, arXiv preprint **arXiv:1705.04241v1 [math.ST]** (2017).

[BKM19] Jose Blanchet, Yang Kang, and Karthyek Murthy, *Robust wasserstein profile inference and applications to machine learning*, arXiv preprint **arXiv:1610.05627v3 [math.ST]** (2019).

[BKS19] Jose Blanchet, Yang Kang, and Nian Si, *Confidence regions in wasserstein distributionally robust estimation*, arXiv preprint **arXiv:1906.01614v1 [math.ST]** (2019).

[BMFR18] Eli Ben-Michael, Avi Feller, and Jesse Rothstein, *The Augmented Synthetic Control Method*, Papers 1811.04170, arXiv.org, November 2018.

[BN13] Andreas Billmeier and Tommaso Nannicini, *Assessing economic liberalization episodes: A synthetic control approach*, The Review of Economics and Statistics **95** (2013), no. 3, 983–1001.

[CGNP13] Eduardo Cavallo, Sebastian Galiani, Ilan Noy, and Juan Pantano, *Catastrophic natural disasters and economic growth*, The Review of Economics and Statistics **95** (2013), no. 5, 1549–1561.

[CMM18] Carlos Carvalho, Ricardo Masini, and Marcelo C. Medeiros, *Arco: An artificial counterfactual approach for high-dimensional panel time-series data*, Journal of Econometrics **207** (2018), no. 2, 352 – 380.

[CWZ17] Victor Chernozhukov, Kaspar Wuthrich, and Yinchu Zhu, *An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls*, Papers 1712.09089, arXiv.org, December 2017.

[Dav80] Robert B. Davies, *Algorithm as 155: The distribution of a linear combination of χ^2 random variables*, Journal of the Royal Statistical Society. Series C (Applied Statistics) **29** (1980), no. 3, 323–333.

[Dav18] Powell David, *Imperfect synthetic controls: Did the massachusetts health care reform save lives?*, Tech. report, 2018.

[DI16] Nikolay Doudchenko and Guido W. Imbens, *Balancing, regression, difference-in-differences and synthetic control methods: A synthesis*, Working Paper 22791, National Bureau of Economic Research, October 2016.

[FP17] Bruno Ferman and Cristine Pinto, *Placebo Tests for Synthetic Controls*, MPRA Paper 78079, University Library of Munich, Germany, April 2017.

[Hal05] A.R. Hall, *Generalized method of moments*, Advanced texts in econometrics, Oxford University Press, 2005.

[Hay11] F. Hayashi, *Econometrics*, Princeton University Press, 2011.

[HS17] Jinyong Hahn and Ruoyao Shi, *Synthetic control and inference*, Econometrics **5** (2017), no. 4, 52.

[IMH61] J. P. IMHOF, *Computing the distribution of quadratic forms in normal variables*, Biometrika **48** (1961), no. 3-4, 419–426.

[Li17] Kathleen T. Li, *Estimating average treatment effects using a modified synthetic control method: Theory and applications*, Tech. report, 2017.

[LTZ09] Huan Liu, Yongqiang Tang, and Hao Helen Zhang, *A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables*, Computational Statistics & Data Analysis **53** (2009), no. 4, 853 – 856.

[Pow16] David Powell, *Synthetic Control Estimation Beyond Case Studies Does the Minimum Wage Reduce Employment?*, Working Papers WR-1142, RAND Corporation, March 2016.

[PY15] Giovanni Peri and Vasil Yasenov, *The labor market effects of a refugee wave: Applying the synthetic control method to the mariel boatlift*, Working Paper 21801, National Bureau of Economic Research, December 2015.

[RSK17] Michael W. Robbins, Jessica Saunders, and Beau Kilmer, *A framework for synthetic control methods with high-dimensional, micro-level data: Evaluating a neighborhood-specific crime intervention*, Journal of the American Statistical Association **112** (2017), no. 517, 109–126.

[SO77] J. Sheil and I. O'Muircheartaigh, *Algorithm as 106: The distribution of non-negative quadratic forms in normal variables*, Journal of the Royal Statistical Society. Series C (Applied Statistics) **26** (1977), no. 1, 92–98.

[SV18] Firpo Sergio and Possebom Vitor, *Synthetic Control Method: Inference, Sensitivity Analysis and Confidence Sets*, Journal of Causal Inference **6** (2018), no. 2, 1–26.

[WHI+15] L. Wong, H. Hong, G. Imbens, F.A. Wolak, and Stanford University. Department of Economics, *Three essays in causal inference*, 2015.

[Xu17] Yiqing Xu, *Generalized synthetic control method: Causal inference with interactive fixed effects models*, Political Analysis **25** (2017), no. 1, 57–76.