Inference in Synthetic Control Methods using the Robust Wasserstein Profile function

Isaac Meza Lopez
Department of IEOR
-University of California at Berkeley
isaac.meza@berkeley.edu

Abstract

A popular method in comparative case studies is the *synthetic control method* (SCM). A problem in this methodology is how to conduct formal inference. This work contributes by using a novel approach similar to Empirical Likelihood (EL), to recover confidence regions, specifically we apply the Robust Wasserstein Profile Inference developed in [BKM19]. 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.

1 Synthetic control method

Synthetic control methods (SCM) are a popular approach in causal inference in comparative studies: [ADH15; BCL$^+$18; ADH10; AG03; PY15; BN13; AI17; CGNP13; AJK$^+$16; RSK17]. 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. There has also been a rich literature extending such methods: [Pow16; Xu17; AL18; DI16; ASS18; BMFR18; Dav18]. This gives an illustration of the importance of the methodology in the causal inference literature in comparative case studies.

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. 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

[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.

Preprint. Under review.

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 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].

2 SCM

The framework is based on the Rubin's potential outcomes setup. Let there be T time periods indexed by $t = 1, \dots, T$ and N sub-populations indexed by $n = 0, 1, \dots, N$ 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. *There exists weights $\beta_n \in [0, 1]$ for $n = 1, \dots, N$ such that*

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

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

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

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

and the overall ATE is

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

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

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

i.e.

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

3 Robust Wasserstein Profile Inference

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

$$\min_{\theta : G(\theta) \leq 0} \mathbb{E}_{P_{\text{TRUE}}} \left[H(X, Y, \theta) \right] \tag{3}$$

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 (3) 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

$$\min_{\theta : G(\theta) \leq 0} \max_{\mathcal{D}_c(P, P_n) \leq \lambda} \mathbb{E}_P \left[H(X, Y, \theta) \right] \tag{4}$$

Here P_n is the empirical measure[2], $\mathcal{D}_c$ is defined to be the Wasserstein distance function[3] with cost c, and δ is called the *distributionally uncertainty size*. We will refer as $\hat{\theta}_n^{DRO}$ to any solution of (4). 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

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

[3] 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.

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

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 (3) 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

$$R_n(\theta) := \inf\{\mathcal{D}_c(P,P_n) \ : \ \mathbb{E}_P\left[h(X,Y,\theta)\right] = 0\} \tag{5}$$

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

Proposition 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.*

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 $\Delta_n(\delta)$.

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 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,*

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

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 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

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

[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.

(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 here is to derive the exact uncertainty set Δ. In [BKS19] a theorem is presented giving the asymptotic normality of underlying DRO estimators, and we reproduce them in the appendix for completeness of exposition. We will use this results in the next section to derive the asymptotics of the SCM estimator.

3.1 Inference via the RWP function

Proposition (1), and an application of Theorem (3) will be the basis to derive an exact confidence region for β. On the other hand, Theorems (6), and (7), will give the asymptotic behaviour of this confidence region.

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

$$\min_{\beta \,:\, ||\beta||_1 = 1 \;\beta_i \geq 0} \mathbb{E}_{P_{T_0}} ||Y - X^\mathsf{T}\beta||^2 \tag{6}$$

Note that the KKT conditions are

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

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

$$h(x,y,\beta,\lambda,\mu,s) = \begin{bmatrix} (y - \beta^\mathsf{T} x)x - \lambda e + \mu \\ 1 - ||\beta||_1 \\ \beta - s^2 \\ \mathrm{diag}(\mu)\,\mathrm{diag}(s) \end{bmatrix} \tag{7}$$

and apply Theorem 3 to this function, to derive our main result.

Theorem 4. *Consider $h(x,y,\beta,\lambda,\mu,s)$ as defined by (7) 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 $A = \sigma^2\Sigma - \lambda^{*2} ee^\mathsf{T} + \lambda^* e\mu^{*\mathsf{T}} + \lambda^*\mu^* e^\mathsf{T} - \mu^*\mu^{*\mathsf{T}}$

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 eigen-descomposition , 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 $\bar{R}(\beta^*) = N^\mathsf{T}\Lambda(\beta^*)N$.

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.

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

$$\Delta_{T_0}(\chi_{1-\alpha}) = \{\beta \mid R_{T_0}(\beta) \leq T_0^{-1}\chi_{1-\alpha}\} \tag{8}$$

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 (2):

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

[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.

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

In order to solve the inequality (8) 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

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

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

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

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

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

We can contrast this procedure with the asymptotic behaviour of the *ERM* estimator and the confidence region given by Theorems 6 and 7.

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

Theorem 5. *Consider the same framework of Theorem (4), 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. Considering the confidence region à la Manski, i.e. by considering the worst-case scenario, 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 $\Sigma B^{-1}\Sigma$

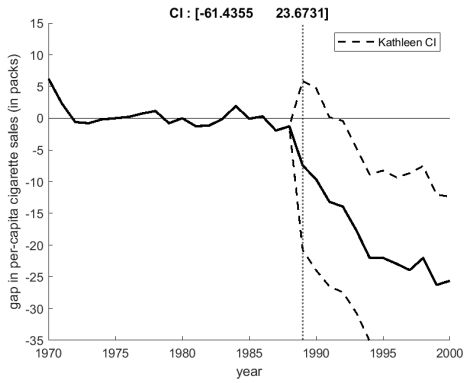
and so the confidence region for the ATE of the SCM is given by

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

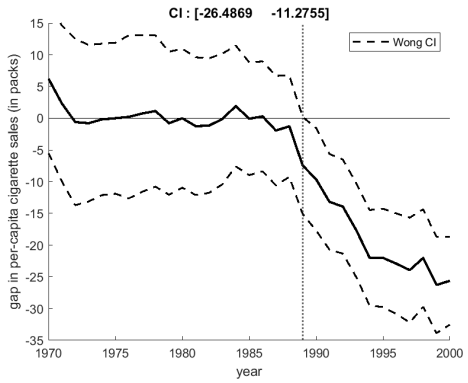
4 Empirical example

In this section we illustrate the method described here with an empirical application, and contrast it with existent methods. 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. Figure 1 contrasts different inference procedures, together with the one presented here. We note that all methods conclude overall significance, excepting Kathleen's [Li17] sub-sampling procedure (Panel a), and the asymptotic Wasserstein CI (Panel d). We first turn our attention to the GMM-derived confidence interval in (Panel b), which is a robustification of the Generalized Method of Moments [WHI+15]. The CI derived with the asymptotics obtained in Theorem (5) (Panel c) 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 c) shows the confidence interval using the Wasserstein profile function; it shows both the 'exact' confidence interval and its asymptotic approximation. (Panel d) only displays the asymptotic Wasserstein confidence interval.

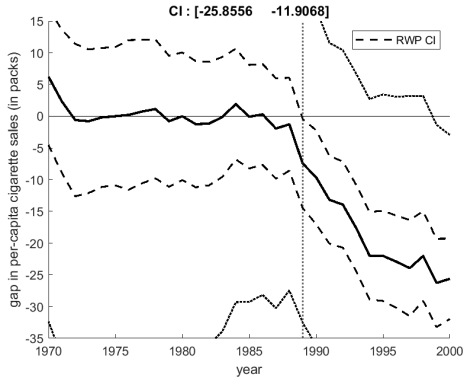
(a) Kathleen sub-sampling



(b) Wong GMM



(c) RWP CI



(d) Asymptotic RWP CI

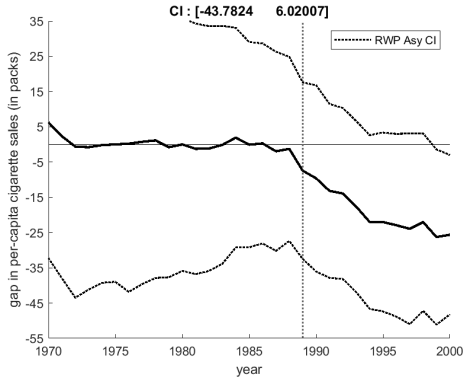


Figure 1: Analytical confidence intervals

5 Conclusion

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 a the linear model 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 data follows a weakly stationary process 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).

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.

References

[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.

[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.

[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.

Appendix

Theorem 6. *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* $\mathbb{E}[D_x h(X, Y, \theta^*) D_x h(X, Y, \theta^*)^\intercal]$ *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 7. *Let* C_n *be a consistent estimator of* $C = \mathbb{E}[D_\theta h\,(X,Y,\theta^*)]$, *and* $\delta(n) = \delta + o(1)$*; then*

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

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

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

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. □

Proof of Theorem 4. To show that the RWP function converges in distribution, we verify the assumptions of Theorem (3) with $h(\cdot)$ defined in (7).

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

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

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,

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

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 (3)

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

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

$$H = [\ \mathcal{Z} \mid \delta_{2N+1}\ ]$$

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. □