cemmap
centre for microdata methods and practice

Inference on average treatment effects in aggregate panel data settings

Victor Chernozhukov
Kaspar Wüthrich
Yinchu Zhu

The Institute for Fiscal Studies
Department of Economics,
UCL

cemmap working paper CWP32/19

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Inference on average treatment effects in aggregate panel data settings*

Victor Chernozhukov† Kaspar Wüthrich‡ Yinchu Zhu§

Abstract

This paper studies inference on treatment effects in aggregate panel data settings with a single treated unit and many control units. We propose new methods for making inference on average treatment effects in settings where both the number of pre-treatment and the number of post-treatment periods are large. We use linear models to approximate the counterfactual mean outcomes in the absence of the treatment. The counterfactuals are estimated using constrained Lasso, an essentially tuning free regression approach that nests difference-in-differences and synthetic control as special cases. We propose a K-fold cross-fitting procedure to remove the bias induced by regularization. To avoid the estimation of the long run variance, we construct a self-normalized t-statistic. The test statistic has an asymptotically pivotal distribution (a student t-distribution with $K-1$ degrees of freedom), which makes our procedure very easy to implement. Our approach has several theoretical advantages. First, it does not rely on any sparsity assumptions. Second, it is fully robust against misspecification of the linear model. Third, it is more efficient than difference-in-means and difference-in-differences estimators. The proposed method demonstrates an excellent performance in simulation experiments, and is taken to a data application, where we re-evaluate the economic consequences of terrorism.

*We are grateful to Yixiao Sun and conference participants at UCL for valuable comments.

†email: vchern@mit.edu

‡email: kwuthrich@ucsd.edu

§email: yzhu6@uoregon.edu

1 Introduction

This paper studies the problem of making inference on treatment effects in aggregate panel data settings with a single treated unit. The treated unit is observed for a number of time periods before and after the intervention occurs. In addition, there are potentially very many untreated units, which serve as controls. Such settings are ubiquitous in applied economic research and there are many different estimation and inference approaches. Examples include difference-in-differences methods (e.g., Ashenfelter and Card, 1985; Card and Krueger, 1994; Bertrand et al., 2004; Athey and Imbens, 2006; Angrist and Pischke, 2008), synthetic control approaches (e.g., Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015; Li, 2017), penalized regression models (e.g., Valero, 2015; Doudchenko and Imbens, 2016; Li and Bell, 2017; Carvalho et al., 2017), and factor, matrix completion and interactive fixed effects models (e.g., Bai, 2003; Pesaran, 2006; Bai, 2009; Hsiao et al., 2012; Kim and Oka, 2014; Gobillon and Magnac, 2016; Chan and Kwok, 2016; Xu, 2017; Athey et al., 2017; Amjad et al., 2017; Li, 2018). Following Chernozhukov et al. (2017), we refer to these methods as counterfactual and synthetic control (CSC) methods.

Here, we consider estimation and inference on average treatment effects in settings where both the number of pre-treatment periods T_0 and the number of post-treatment periods T_1 are large. We approximate the counterfactual outcome of the treated unit in the absence of the treatment, Y_{0t}^N, using a linear model:

$$Y_{0t}^N = \mu + \sum_{i=1}^{N} w_i Y_{it}^N + u_t, \quad E(u_t) = 0, \quad t = 1, \ldots, T_0 + T_1. \tag{1}$$

We are interested in testing hypotheses about the average treatment effect (ATE):

$$\tau = \frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \alpha_t,$$

where $\alpha_t = Y_{0t}^I - Y_{0t}^0$ is the per-period treatment effect, which is equal to the difference between the potential outcome with the treatment, Y_{0t}^I, and Y_{0t}^N. We propose estimating w using constrained Lasso (e.g., Raskutti et al., 2011; Chernozhukov et al., 2017), which imposes an ℓ_1 constraint on the weight vector w. Constrained Lasso is an essentially tuning regression approach which nests two of the most popular approaches for estimating counterfactuals in aggregate panel settings, difference-in-differences and synthetic control, as special cases.

We develop a K-fold cross fitting scheme for bias-correction to obtain a consistent and asymptotically normal estimator. The key assumption underlying our procedure is stationarity of the data. Consequently, if the data are not stationary, we first pre-process the data to make them stationary before applying our method.

Inference on τ is based on a t-statistic that exploits a self-normalization structure and thereby avoids the tricky issue of estimating the long-run variance. The resulting test

statistic has an asymptotically pivotal distribution (a student t-distribution with $K-1$ degrees of freedom), which makes our inference procedure very easy to implement. The construction of our test statistic is inspired by Ibragimov and Müller (2010).

The proposed new method has several theoretical advantages. First, while we allow for high-dimensional covariates, we do not impose any sparsity assumption on w. As a result, the proposed procedure is robust to the lack of sparsity. Second, the validity of the proposal does not rely on the correct specification of the linear model. This is a very appealing feature of our work because, in practice, misspecification is quite difficult to rule out. Third, compared to difference-in-means and difference-in-differences estimators, our method is more efficient in terms of asymptotic variance.

In deriving the theoretical results, we obtain a result that may be of independent interest. We provide a characterization of constrained Lasso under misspecification. Specifically, we show that the constrained Lasso estimator is consistent in ℓ_2 for the pseudo-true parameter value, which is defined as the minimizer of the population loss subject to the ℓ_1 constraint.

The proposed method demonstrates an excellent performance in simulation experiments, and is taken to a data application, where we re-visit the analysis of the economic consequences of terrorism in Abadie and Gardeazabal (2003).

1.1 Related literature

Here, we discuss the relationship of our approach and existing CSC methods based on linear models. References to inference procedures for other types of CSC approaches are provided at the beginning of the introduction.

Our approach is most closely related to a very recent literature which proposes asymptotic inference theory for settings where both T_0 and T_1 are large. Focusing on the expected treatment effect $E(\alpha_t)$, Li and Bell (2017) derive the asymptotic distribution of the least squares estimator proposed by Hsiao et al. (2012).[1] Moreover, they propose to use Lasso to select the relevant control units, but do not provide formal theory. In related work, Li (2017) studies inference on $E(\alpha_t)$ in synthetic control settings. In this paper, we focus on a different target, the ATE, $\frac{1}{T_1}\sum_{t=T_0+1}^{T_0+T_1}\alpha_t$, which allows us to avoid strong assumptions on $\{\alpha_t\}$. Carvalho et al. (2017) study inference on the ATE in settings where the parameters are estimated using Lasso. Their approach relies sparsity of w. By contrast, we develop a cross fitting scheme for bias-correction that allows us to completely avoid any sparsity assumptions. Another key difference to the literature (e.g., Li and Bell, 2017; Carvalho et al., 2017), which typically relies on Newey-West-type variance estimators, is that we rely on a self-normalized test statistic. This allows us to completely avoid the estimation of the long run variance, which we found to be essential for achieving a good performance in small sample settings.

[1] Note that the linear models in Li and Bell (2017) and Hsiao et al. (2012) are derived from a factor structure.

Another part of the literature focuses on finite population inference approaches. These approaches assume that the potential outcomes are fixed but unknown, which justifies the use of permutation/randomization tests for testing sharp null hypotheses about the whole treatment effect trajectory $\{\alpha_t\}_{t=T_0+1}^{T_0+T_1}$. This approach was introduced in the context of synthetic control methods (Abadie et al., 2010), but can be applied to a much broader class of linear methods, including difference-in-differences and penalized regression models (e.g., Doudchenko and Imbens, 2016). We refer to Firpo and Possebom (2017) and Ferman and Pinto (2017) for a discussion of these approaches. Our proposal differs from these finite population approaches in that we study the problem of making inferences on the ATE in a super-population setting where the number of time periods and the number of potential control units are large.

Finally, Chernozhukov et al. (2017) propose a generic conformal inference approach for testing sharp null hypotheses about $\{\alpha_t\}_{t=T_0+1}^{T_0+T_1}$. Their method can be combined with linear models such as constrained Lasso. Here, we focus on a different target, the ATE, and rely on asymptotic approximations instead of permutation distributions for making inference.

1.2 Plan of the paper

We introduce some frequently used notations. For $q \geq 1$, the ℓ_q-norm of a vector is denoted by $\|\cdot\|_q$. We use $\|\cdot\|_0$ to denote the number of nonzero entries of a vector; $\|\cdot\|_\infty$ is used to denote the maximal absolute value of entries of a vector. For a matrix A, we use the notation $\|A\|_\infty = \|\text{vec}(A)\|_\infty$, where $\text{vec}(A)$ is the column-wise vectorization of A. We also use the notation $a \lesssim b$ to denote $a \leq cb$ for some constant $c > 0$ that does not depend on the sample size. We also use the notation $a \asymp b$ to denote $a \lesssim b$ and $b \lesssim b$. For a set A, $|A|$ denotes the cardinality of A.

Section 2 introduces the setup and our methodology. In Section 3, we derive theoretical properties of our method. Section 4 provides simulation evidence on the finite sample performance of our procedure. In Section 5, we use our method to re-analyze the economics consequences of terrorism. Section 6 concludes. All proofs are collected in the appendix.

2 Methodology

2.1 Setup

Consider an aggregate panel data setting with $N+1$ units and T time periods. Unit $i = 0$ is the treated unit and units $i = 1, \ldots, N$ are the control units. The treated unit is untreated for the first T_0 periods, and treated during the remaining $T - T_0 = T_1$ periods. The control units remain untreated for all T periods. We observe $\{(Y_{it}, Z_{it})\}$, where Y_{it} is the outcome of interest and Z_{it} contains additional observable characteristics.

Our analysis is developed within the potential (latent) outcome framework (Neyman, 1923; Rubin, 1974). Potential outcomes with and without the treatment are denoted as Y_{it}^I and Y_{it}^N. Observed outcomes are related to potential outcomes as $Y_{it} = D_{it}Y_{it}^I + (1-D_{it})Y_{it}^N$, where the treatment indicator can be written as $D_{it} = \mathbf{1}\{t > T_0, i = 0\}$. Our object of interest is the ATE in the post treatment period:

$$\tau = \frac{1}{T_1} \sum_{t=T_0+1}^{T} \alpha_t, \tag{2}$$

where $\alpha_t := \alpha_{0t} = Y_{0t}^I - Y_{0t}^N$ is the per-period treatment effect. Specifically, we consider the problem of testing hypotheses of the form

$$H_0 : \tau = \tau_0. \tag{3}$$

Because the counterfactual outcome Y_{0t}^N is fundamentally unobserved for $t > T_0$, α_t is not identified without additional assumptions. To overcome this identification problem, we impose a linear model for $Y_t^N := Y_{0t}^N$, which we write compactly as

$$Y_t^N = X_t'w + u_t,$$

where $X_t \in \mathbb{R}^p$ is a vector of transformations of $\left(Y_{1t}^N, \ldots, Y_{Nt}^N, Z_{0t}, \ldots, Z_{Nt}\right)$ as in model (1) and $w \in \mathbb{R}^p$ is an unknown parameter. To improve the accuracy of predicting the mean of Y_t^N, p is allowed to be large, even larger than T_0. For concreteness, we consider the constrained Lasso estimator (e.g., Raskutti et al., 2011; Chernozhukov et al., 2017), which restricts w to be in a subset of an ℓ_1-ball of bounded radius Q, $\|w\|_1 \leq Q$:

$$\hat{w} = \arg\min_w \frac{1}{T_0} \sum_{t=1}^{T_0} (Y_t - X_t'w)^2 \quad \text{s.t.} \quad \|w\|_1 \leq Q. \tag{4}$$

One can also include an unregularized intercept in the above estimator, but this is numerically equivalent to running the above estimator with demeaned data. Hence, for our theoretical analysis, we focus on the estimator in (4).

2.2 Bias-corrected estimation and inference

The natural estimator for τ would be $T_1^{-1} \sum_{t=T_0+1}^{T} (Y_t - X_t'\hat{w})$ with $\hat{w}$ defined in (4). However, this estimator is biased due to the bias in $\hat{w}$. To remove the bias, we propose a K-fold cross-fitting procedure. Throughout this paper, K is assumed to be fixed. We partition the pre-treatment period into K consecutive pieces: $H_1 \bigcup H_2 \bigcup \cdots \bigcup H_K = \{1, \ldots, T_0\}$. Define $r = \lfloor T_0/K \rfloor$ and let $H_k = \{(k-1)r+1, \ldots, kr\}$ for $k \leq K-1$ and $H_K = \{(K-1)r+1, \ldots, T_0\}$. For notational simplicity, we assuem that T_0/K is an integer. The constrained Lasso estimator using data in $H_{(-k)} := \{1, \ldots, T_0\} \backslash H_k$ is given by

$$\hat{w}_{(k)} = \arg\min_w \sum_{t \in H_{(-k)}} (Y_t - X_t'w)^2 \qquad s.t. \qquad \|w\|_1 \leq Q.$$

Define

$$\hat{\tau}_k = \frac{1}{T_1} \sum_{t=T_0+1}^{T} (Y_t - X_t' \hat{w}_{(k)}) - \frac{1}{|H_k|} \sum_{t \in H_k} (Y_t - X_t' \hat{w}_{(k)}).$$

The estimator for the treatment effect is

$$\hat{\tau} = \frac{1}{K} \sum_{k=1}^{K} \hat{\tau}_k. \tag{5}$$

To avoid estimating σ, we construct a test statistic that is scale-free. The idea is to form a ratio in which the numerator and denominator are both scaled by the long-run standard deviation. Specifically, we propose to use a t-statistic based on $\{\hat{\tau}_k\}$:

$$\mathbb{T}_K = \frac{\sqrt{K}\,(\hat{\tau} - \tau)}{\hat{\sigma}_{\hat{\tau}}}, \tag{6}$$

where

$$\hat{\sigma}_{\hat{\tau}} = \sqrt{1 + \frac{T_0}{T_1}} \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} (\hat{\tau}_k - \hat{\tau})^2}.$$

The construction of the test statistic $\mathbb{T}_K$ is inspired by Ibragimov and Müller (2010).

3 Theory

3.1 Validity under correction specification

In this subsection, we derive the first theoretical result of the paper. Under correct specification, the proposed estimator in (5) is shown to be $\sqrt{T_0}$-consistent and asymptotically normal and the test statistic $\mathbb{T}_K$ in (6) is shown to have a student t-distribution with $K-1$ degrees of freedom.

We summarize our model in the following assumption.

Assumption 1 (Model)**.**

$$Y_t^N = X_t' w + u_t, \tag{7}$$

where $E(u_t) = 0$, $E\left(X_t u_t\right) = 0$ *and* $\|w\|_1 \leq Q$ *with* $Q = O(1)$.

Remark 1. The constrained Lasso model in Assumption 1 nests two of the most popular approaches for modeling counterfactuals based on linear models, difference-in-differences and canonical synthetic controls, as special cases. (As argued before, we can easily include an intercept μ.) The classical difference-in-differences model imposes the following model for the counterfactual in the absence of the treatment (e.g., Doudchenko and Imbens, 2016):

$$Y_t^N = \mu + \frac{1}{N} \sum_{i=1}^{N} Y_{it}^N + u_t. \tag{8}$$

This model is nested by Assumption 1 by setting $w_i = 1/N$ for all $i = 1, \ldots, N$, $X_t = \left(Y_{1t}^N, \ldots, Y_{Nt}^N\right)'$, and $Q = 1$. The canonical synthetic control model (Abadie et al., 2010, 2015; Doudchenko and Imbens, 2016) assumes that

$$Y_t^N = \sum_{i=1}^N w_i Y_{it}^N + u_t, \quad w_i > 0, \quad \sum_{i=1}^N w_i = 1 \tag{9}$$

This model is nested by setting $\mu = 0$, $w_i > 0$, $X_t = \left(Y_{1t}^N, \ldots, Y_{Nt}^N\right)'$, and $Q = 1$.

To study the asymptotic properties of $\hat{\tau}$, we first establish the ℓ_2-consistency of constrained Lasso for estimating $w_{(k)}$. For this, we impose the following assumptions.

Assumption 2. *Let K be fixed. Suppose that the following conditions hold for $1 \leq k \leq K$:*

1. *Let $\Sigma_{(-k)} = |H_{(-k)}|^{-1} \sum_{t \in H_{(-k)}} E(X_t X_t')$. There exists a constant $\kappa > 0$ such that $\min_{1 \leq k \leq K} \lambda_{\min}(\Sigma_{(-k)}) \geq \kappa$.*
2. *$\| |H_{(-k)}|^{-1} \sum_{t \in H_{(-k)}} (X_t X_t' - E(X_t X_t')) \|_\infty = o_P(1)$ and $\| |H_{(-k)}|^{-1} \sum_{t \in H_{(-k)}} X_t u_t \|_\infty = o_P(1)$.*

The eigenvalues of $\Sigma_{(-k)}$ are assumed to be bounded away from zero to achieve identification of w. The second condition in Assumption 2 holds under weak serial dependence, mild conditions on the tail of the distribution of the variables and conditions on p. For example, when entries of X_t and u_t are sub-Gaussian, we can allow for $\log p = o(\sqrt{T_0})$; when entries of X_t and u_t have bounded rth moment for $r > 2$, then we can typically allow for $p = o(T_0^{r/4})$.

The next theorem presents our first main result.

Theorem 1. *Let Assumptions 1 and 2 hold. Then $\|\hat{w}_{(k)} - w\|_2 = o_P(1)$. In particular,*

$$\left\|\hat{w}_{(k)} - w\right\|_2^2 \leq \frac{4 \| |H_{(-k)}|^{-1} \sum_{t \in H_{(-k)}} (X_t X_t' - E(X_t X_t')) \|_\infty Q^2 + 4 \| |H_{(-k)}|^{-1} \sum_{t \in H_{(-k)}} X_t u_t \|_\infty Q}{\kappa}.$$

Remark 2. Under Gaussianity, one could sharpen the bound in Theorem 1; see Theorem 1 of Raskutti et al. (2010) which has a slightly different rate. However, according to Rudelson and Zhou (2013), the argument in Raskutti et al. (2010) cannot be extended to the non-Gaussian case because their argument exploits the Gaussianity through Gordon's Theorem, which is not available beyond Gaussian settings.

To establish the asymptotic distribution, we impose the following stationarity assumption.

Assumption 3 (Stationarity). *$E(X_t)$ does not depend on t and $\{u_t\}_{t=1}^T$ is covariance-stationary.*

We would like to emphasize that stationarity as stated in Assumption 3 is not just a technical regularity condition but is one of the key assumptions underlying our results. If the data are not stationary in that the mean of X_t depends on t, one has to pre-process the data to make it stationarity before applying our methodology. For example, in the application in Section 5, since the GDP data is non-stationary, we de-trend the data before applying our method. We would like to note that stationarity is a common fundamental assumption imposed for high-dimensional models in the literature. For example, Carvalho et al. (2017) assume that the entire (X_t, Y_t^N) is fourth-order stationary. For non-stationary data, only highly parametrized models are considered; for example, Li (2018) imposes a factor structure for the outcome in a panel data setting and allows the factor to follow an exact unit-root process.

By simple algebra, Assumption 3 gives us the following observation.

Lemma 1. *Let Assumptions 1 and 3 hold. Then for any $1 \leq k \leq K$,*

$$\hat{\tau}_k - \tau = \left(T_1^{-1} \sum_{t=T_0+1}^{T} u_t - \frac{1}{|H_k|} \sum_{t \in H_k} u_t \right) + \left(\frac{1}{|H_k|} \sum_{t \in H_k} \tilde{X}_t - T_1^{-1} \sum_{t=T_0+1}^{T} \tilde{X}_t \right)' \Delta_{(k)},$$

where $\tilde{X}_t = X_t - E(X_t)$ and $\Delta_{(k)} = \hat{w}_{(k)} - w$.

The following weak conditions guarantee that the second term in Lemma 1 is negligible. The idea is that under weak dependence, $\Delta_{(k)}$ is approximately independent of data in $H_k \bigcup \{T_0 + 1, ..., T\}$. Hence, the ℓ_2-consistency established in Theorem 1 can be used to bound the second term in Lemma 1.

Assumption 4. *Suppose the following conditions hold:*

1. *There exists a constant $\kappa_1 > 0$ such that for any $A \subseteq \{1, \ldots, T_0\}$, the largest eigenvalue of*

$$E\left[|A|^{-1} \left(\sum_{t \in A} \tilde{X}_t \right) \left(\sum_{t \in A} \tilde{X}_t \right)' \right]$$

   is bounded above by κ_1.

2. *There exists a sequence $\rho_{T_0} > 0$ such that $P(\max_{1 \leq t \leq T_0} \|\tilde{X}_t\|_\infty \leq \rho_{T_0}) \to 1$.*

3. *The data $\{(X_t, u_t)\}_{t=1}^T$ is β-mixing with coefficient satisfying $\beta_{\text{mix}}(\gamma_T) \to 0$ for some sequence γ_T satisfying $0 < \gamma_T < T_0/(K+1)$ and $\rho_T \gamma_T = o(\min\{\sqrt{T_0}, \sqrt{T_1}\})$.*

The weak dependence is stated in terms of β-mixing, which holds for a large class of stochastic processes. The bound on $\tilde{X}_t$ is allowed to grow but no faster than $\min\{\sqrt{T_0}, \sqrt{T_1}\}$. We are now in the position to state our second main result.

Theorem 2. *Let Assumptions 1, 3, and 4 hold. Suppose that* $\|w\|_1 = O(1)$ *and* $Q = O(1)$. *Also assume* $T_1 = O(T_0)$. *Then we have*

$$\max_{1\le k\le K}\left|\hat{\tau}_k - \tau - \left(T_1^{-1}\sum_{t=T_0+1}^{T} u_t - \frac{1}{|H_k|}\sum_{t\in H_k} u_t\right)\right| = o_P\left(\frac{1}{\min\{\sqrt{T_0},\sqrt{T_1}\}}\right) + O_P\left(\frac{\max_{1\le k\le K}\|\Delta_{(k)}\|_2}{\min\{\sqrt{T_0},\sqrt{T_1}\}}\right).$$

By Theorem 2, the asymptotic normality follows once we show $\max_{1\le k\le K}\|\Delta_{(k)}\|_2 = o_P(1)$. Assumption 2 provides sufficient conditions for constrained Lasso. Theorem 1 and Theorem 2 then imply the following key result.

Corollary 1. *Let Assumptions 1, 2, 3, and 4 hold. Suppose that* $T_0/T_1 \to c_0$ *for some* $c_0 \in [0,\infty)$. *Then*

$$\sqrt{T_0}\,(\hat{\tau}_k - \tau) = \sqrt{T_0/T_1}\left(T_1^{-1/2}\sum_{t=T_0+1}^{T} u_t\right) - \sqrt{T_0/|H_k|}\left(\frac{1}{|H_k|^{-1/2}}\sum_{t\in H_k} u_t\right) + o_P(1).$$

Moreover, if $\{u_t\}_{t=1}^T$ *satisfies* $\max_{1\le t\le T_0} E|u_t|^r = O(1)$ *and* $\beta_{\text{mix}}(i) \lesssim i^{-\eta}$ *for some constants* $r > 2$ *and* $\eta > r/(r-2)$, *then*

$$\sqrt{T_0}\begin{pmatrix}\hat{\tau}_1 - \tau\\ \hat{\tau}_2 - \tau\\ \vdots\\ \hat{\tau}_K - \tau\end{pmatrix} \xrightarrow{d} \begin{pmatrix}\sqrt{c_0}\xi_0 - \sqrt{K}\xi_1\\ \sqrt{c_0}\xi_0 - \sqrt{K}\xi_2\\ \vdots\\ \sqrt{c_0}\xi_0 - \sqrt{K}\xi_K\end{pmatrix}\sigma,$$

where $\xi_0,\ldots,\xi_K$ *are independent* $N(0,1)$ *random variables and* $\sigma^2 = \lim_{T\to\infty} E\left(T^{-1/2}\sum_{t=1}^{T} u_t\right)^2$.

The asymptotic normality in Corollary 1 follows by the usual CLT for dependent processes, e.g., Theorem 5.20 of White (2014).

Notice that we allow for the case $T_1 \gg T_0$, i.e., $c_0 = 0$. In this scenario, the variance comes from the errors in the post-treatment periods. Moreover, we include the case in which $T_1 \asymp T_0$. This is a relevant scenario in many applications.

The next theorem establishes the asymptotic distribution of our estimator and test statistic.

Theorem 3. *Let all the assumptions in Corollary 1 hold. Then (i)*

$$\sqrt{T_0}(\hat{\tau} - \tau) \xrightarrow{d} N(0, (1 + c_0)\sigma^2)$$

and (ii)

$$\mathbb{T}_K \xrightarrow{d} t_{K-1}$$

where the random variable t_{K-1} *has a standard t-distribution with* $K-1$ *degrees of freedom.*

The first part of Theorem 3 establishes the asymptotic normality of our ATE estimator. Making inference directly based on this result would require estimation of the long run variance σ^2, which can be tricky in small sample settings. We therefore use the self-normalized test statistic $\mathbb{T}_K$, which allows us to completely avoid estimation of σ^2. The second part of the Theorem 3 demonstrates that $\mathbb{T}_K$ has an asymptotically pivotal student t-distribution with $K-1$ degrees of freedom. This result is very useful from a practical perspective, as one does not have to simulate non-standard critical values. Theorem 3 further suggests the following $1-\alpha$ confidence interval for τ:

$$CI_K(1-\alpha) = \left[\hat{\tau} - t_{K-1}(1-\alpha/2)\frac{\hat{\sigma}_{\hat{\tau}}}{\sqrt{K}},\ \hat{\tau} + t_{K-1}(1-\alpha/2)\frac{\hat{\sigma}_{\hat{\tau}}}{\sqrt{K}}\right],$$

where $t_{K-1}(1-\alpha/2)$ denotes the $(1-\alpha/2)$-quantile of t-distribution with $K-1$ degrees of freedom.

Notice that $\mathbb{T}_K$ and the limiting distribution depend on the choice of K. The choice of K cannot be avoided since it is inherent in the cross-fitting procedure that we employ to obtain asymptotic normality. To shed some light on the choice of K, we analyze the expected width of the confidence interval.

By Corollary 1, we have that $\sqrt{T_0}|CI_K(1-\alpha)| \xrightarrow{d} \zeta_{K-1}$, where

$$E\zeta_{K-1} = 2\sigma t_{K-1}(1-\alpha/2)\sqrt{\frac{1+c_0}{K-1}}E\left(\sqrt{\sum_{k=1}^{K}(\xi_k - \bar{\xi})^2}\right),$$

where $\bar{\xi} = K^{-1}\sum_{k=1}^{K}\xi_k$, and $\{\xi_k\}_{k=1}^{K}$ are i.i.d. standard normal random variables. Notice that $\sum_{k=1}^{K}(\xi_k - \bar{\xi})^2$ has a χ^2 distribution with $K-1$ degrees of freedom and thus,

$$E\left(\sqrt{\sum_{k=1}^{K}(\xi_k - \bar{\xi})^2}\right) = \sqrt{2}\frac{\Gamma(K/2)}{\Gamma((K-1)/2)}. \tag{10}$$

Using (10), we can rewrite the expected length of the confidence interval as

$$E\zeta_{K-1} = C \cdot t_{K-1}(1-\alpha/2)\sqrt{\frac{1}{K-1}}\frac{\Gamma(K/2)}{\Gamma((K-1)/2)},$$

where the constant $C = 2\sqrt{2}\sqrt{1+c_0}\sigma$ does not depend on K. Figure 1 plots $E\zeta_{K-1}$ as a function of K, where we set $T_0 = 18$, $T_1 = 25$ (as in our application), $\alpha = 0.1$, and $\sigma = 1$. We can see that the expected length is strictly decreasing in K. Increasing K from 2 to 3 and from 3 to 4 leads to a substantial reductions in the expected length of the confidence intervals, while increasing K beyond 4 or 5 does not reduce the length much further. In practice, the choice of K is subject to a trade-off between the expected length of the confidence interval and its finite sample coverage properties. Choosing K large will lead to shorter confidence intervals, but may impact the accuracy of the t-approximation of $\mathbb{T}_K$ in Theorem 3, which can lead to undercoverage. In Section 4, we investigate the choice of K based on empirical Monte Carlo simulations that are calibrated to match several features of our empirical application.

3.2 Validity under misspecification specification

Assumption 1 might not always hold in practice. For stationary data, while it is true that $E(X_t u_t) = 0$ for $u_t = Y_t - X_t'w$ with $w = [E(X_t X_t')]^{-1}E(X_t Y_t)$, there is no guarantee that $\|w\|_1 \leq Q$. Therefore, a natural question is whether the proposed method is still valid in this situation. Here, we show that the answer is affirmative.

We first provide some intuition. Let $w_* = \arg\min_{\|v\|_1 \leq Q} E(Y_t - X_t'v)^2$ and $u_{*,t} = Y_t - X_t'w_*$. We shall show that under weak conditions, we have $\|\hat{w}_{(k)} - w_*\|_2 = o_P(1)$ and

$$\hat{\tau} - \tau = T_1^{-1} \sum_{t=T_0+1}^{T} u_{*,t} - T_0^{-1} \sum_{t=1}^{T_0} u_{*,t} + o_P(T_0^{-1/2}).$$

Since w_* might not equal to $[E(X_t X_t')]^{-1}E(X_t Y_t)$, there is no guarantee that $E(u_{*,t}) = 0$. However, as long as $\{u_{*,t}\}_{t=1}^T$ is covariance-stationary, we should still expect zero mean and asymptotic normality for $T_1^{-1}\sum_{t=T_0+1}^{T} u_{*,t} - T_0^{-1}\sum_{t=1}^{T_0} u_{*,t}$. Next, we provide regularity conditions to formalize this intuition.

Assumption 5. *Suppose that $\{(X_t, Y_t^N)\}_{t=1}^T$ is covariance-stationary and satisfies the following conditions*
(1) $\|(\hat{\mu}_{(-k)} - \mu) - (\hat{\Sigma}_{(-k)} - \Sigma)w_*\|_\infty = o_P(1)$, where $\mu = EX_t Y_t^N$, $\Sigma = EX_t X_t'$, $\hat{\mu}_{(-k)} = |H_{(-k)}|^{-1}\sum_{t\in H_{(k)}} X_t Y_t^N$ and $\hat{\Sigma}_{(-k)} = |H_{(-k)}|^{-1}\sum_{t\in H_{(k)}} X_t X_t'$.
(2) $\|\hat{\Sigma}_{(-k)} - \Sigma\|_\infty = o_P(1)$ and $\lambda_{\min}(\Sigma) \geq c$.

Assumption 5 serves the role of Assumption 2 in that it is essentially a law of large numbers uniformly across entries of $X_t Y_t^N$ and $X_t X_t'$. Notice that Assumption 5 directly states a condition on $X_t Y_t^N$ instead of $X_t u_t$. Under Assumption 5, we have the following consistency result under misspecification.

Theorem 4. *Let Assumption 5 hold. Assume that $Q = O(1)$. Then $\|\hat{w}_{(k)} - w_*\|_2 = o_P(1)$. In particular,*

$$\|\hat{w}_{(k)} - w_*\|_2^2 \leq \frac{4\|\hat{\Sigma} - \Sigma\|_\infty Q^2 + 2\|\xi\|_\infty Q}{c}.$$

Next, we note that all the arguments in Lemma 1 and Theorem 2 still hold with u_t replaced by $u_{*,t}$. Therefore, we can still have a result analogous to Corollary 1 and Theorem 3.

Corollary 2. *Let Assumptions 4 and 5 hold. Suppose that $Q = O(1)$ and $T_0/T_1 \to c_0$ for some $c_0 \in [0, \infty)$. Then*

$$\sqrt{T_0}\,(\hat{\tau}_k - \tau) = \sqrt{T_0/T_1}\left(T_1^{-1/2}\sum_{t=T_0+1}^{T} \tilde{u}_t\right) - \sqrt{T_0/|H_k|}\left(\frac{1}{|H_k|^{-1/2}}\sum_{t\in H_k} \tilde{u}_t\right) + o_P(1),$$

where $\tilde{u}_t = u_{*,t} - E(u_{*,t})$. *Moreover, if* $\max_{1\le t\le T_0} E|\tilde{u}_t|^r = O(1)$ *and* $\beta_{\text{mix}}(i) \lesssim i^{-\eta}$ *for some constants* $r > 2$ *and* $\eta > r/(r-2)$, *then*

$$\sqrt{T_0}\begin{pmatrix}\hat{\tau}_1 - \tau \\ \hat{\tau}_2 - \tau \\ \vdots \\ \hat{\tau}_K - \tau\end{pmatrix} \xrightarrow{d} \begin{pmatrix}\sqrt{c_0}\xi_0 - \sqrt{K}\xi_1 \\ \sqrt{c_0}\xi_0 - \sqrt{K}\xi_2 \\ \vdots \\ \sqrt{c_0}\xi_0 - \sqrt{K}\xi_K\end{pmatrix}\sigma_*,$$

where $\xi_0, \ldots, \xi_K$ *are independent* $N(0,1)$ *random variables and* $\sigma_*^2 = \lim_{T\to\infty} E\left(T^{-1/2}\sum_{t=1}^T \tilde{u}_t\right)^2$.

Theorem 5. *Let all the assumptions in Corollary 2 hold. Then (i)*

$$\sqrt{T_0}(\hat{\tau} - \tau) \xrightarrow{d} N(0, (1+c_0)\sigma_*^2)$$

and (ii)

$$\mathbb{T}_K \xrightarrow{d} t_{K-1}.$$

Theorem 5 demonstrates that, under stationarity, our inference procedure is fully robust against misspecification of the linear model in Assumption 1.

3.3 Efficiency

In this subsection, we consider the efficiency of our estimator $\hat{\tau}$ defined in (5). Since we assume that $E(Y_t^N)$ does not depend on t, a natural estimator is the difference-in-means

$$\tilde{\tau} = \frac{1}{T_1}\sum_{t=T_0+1}^{T} Y_t - \frac{1}{T_0}\sum_{t=1}^{T_0} Y_t. \tag{11}$$

The difference-in-means estimator (11) is similar in spirit to typical treatment effect estimators used in randomized experiments, where researchers estimate treatment effects by comparing the averages of the treatment and the control group. It is well-known that, in the context of randomized experiments, one can use regression adjustments to improve the efficiency of the simple difference-in-means estimator. The classical argument is based on the analysis of variance and dates back to Neyman (1923). This has recently been considered for growing number of covariates in Lei and Ding (2018); Bloniarz et al. (2016). In Lei and Ding (2018), an OLS estimator is used for each group and it is required that $p \log p = o(n)$; in Bloniarz et al. (2016), the linear model is assumed to be correctly specified and a sparsity requirement is imposed.

The next theorem establishes a similar result in our context.

Theorem 6. *Suppose that* $\{(Y_t^N, X_t)\}_{t=1}^T$ *is i.i.d with* $E(X_t) = 0$, $E(Y_t^N) = 0$ *and* $E|Y_t^N|^{2+\delta} < \infty$ *for some* $\delta > 0$. *Assume that* $T_0/T_1 \to c_0$ *for some* $c_0 \in [0, \infty)$. *Then*

$$\sqrt{T_0}(\tilde{\tau} - \tau) \xrightarrow{d} N(0, (1+c_0)\sigma_0^2),$$

where $\sigma_0^2 = E(Y_t^N)^2$. *Moreover,* $\sigma_0 \ge \sigma_*$, *where* σ_* *is defined in Theorem 5.*

When the covariates and outcome variable do not have mean zero, we can add an unpenalized intercept to the estimation of w. Since doing so is equivalent to applying existing methods to the demeaned data, the assumption of $E(X_t) = 0$ and $E(Y_t^N) = 0$ does not lose much generality. Theorem 6 shows that, asymptotically, the proposed estimator $\hat{\tau}$ is at least as efficient as the difference-in-means estimator $\tilde{\tau}$, irrespective of whether or not the linear model in Assumption 1 is correctly specified.

A similar efficient result can be derived with respect to difference-in-difference estimators. Consider the estimator

$$\tilde{\tau} = \frac{1}{T_1} \sum_{t=T_0+1}^{T} (Y_t - X_t'\beta) - \frac{1}{T_0} \sum_{t=1}^{T_0} (Y_t - X_t'\beta), \tag{12}$$

where $\beta \in \mathbb{R}^p$ is a pre-specified vector. In the difference-in-differences model described in Remark 1, β would be the vector with each entry equal to $1/p$. By a similar argument as in Theorem 6, one can show that the estimator in (12) is asymptotically normal with an asymptotic variance larger than or equal to σ_*^2.

4 Monte Carlo simulations

4.1 Empirical Monte Carlo simulations

Our first simulation study is based on the empirical application in Section 5. We will use these simulations to inform our choice of K in the application. We set $T_0 = 18$, $T_1 = 25$, and $N = 16$ as in the empirical application. The potential outcomes in the absence of the treatment are modeled as

$$Y_t^N = \mu + \sum_{i=1}^{N} w_i Y_{it}^N + u_t, \quad t = 1, \ldots, T.$$

Here, we set μ and w equal to the corresponding estimates based on the pre-treatment period in the application and generate $\{u_t\}$ according to an AR(1) model fitted to the estimated residuals, where we draw innovations from the empirical distribution of the AR(1) residuals. To generate the control outcomes, we first fit separate AR(1) models to $\{Y_{it}^N\}_{t=1}^T$ for all $i = 1, \ldots, N$. Let $\{\hat{\epsilon}_t\}_{t=1}^T$ denote the corresponding residuals, where $\hat{\epsilon}_t = (\hat{\epsilon}_{1t}, \ldots, \hat{\epsilon}_{Nt})$. The control outcomes are simulated according to the estimated AR(1) models, where we draw the innovations for all units jointly from the empirical distribution of $\{\hat{\epsilon}_t\}_{t=1}^T$ to preserve the cross-sectional dependence. The treatment effects are generated as $\alpha_t = \tau_0 + \tilde{\xi}_t$, where $\tilde{\xi}_t = \xi_t - 1/T_1 \sum_{t=T_0+1}^{T} \xi_t$ and $\xi_t \overset{iid}{\sim} N(0, \sigma_{\hat{u}}^2)$ such that the true effect is exactly equal to τ_0 and $\sigma_{\hat{u}}^2$ is the sample variance of the residuals. We do not redraw the treatment effect across simulations and set τ_0 equal to the difference-in-means estimate in the application. We consider $K \in \{2, 3, 4, 5\}$ and compare our method to the Lasso-based

approach for estimating "Artificial Counterfactuals" (ArCo) proposed by Carvalho et al. (2017). The nominal coverage is equal to $1 - \alpha = 0.9$ for all simulations.

Table 1 shows the empirical coverage and average length of the confidence intervals. The simulation results are consistent with our theory. For $K = 2$, the empirical coverage is very close to the nominal level, while the confidence intervals are rather wide on average. Choosing $K = 3$ yields much shorter confidence intervals that still exhibit good coverage properties. Increasing K to 4 and 5 further reduces the average length of the confidence intervals. However, this reduction is accompanied by a substantial deterioration of the finite sample coverage properties. ArCo yields very short confidence intervals that exhibits substantial undercoverage. In view of these empirical Monte Carlo simulations, we choose $K = 3$ in our application.

4.2 Additional simulations

Here, we present additional simulation evidence. Our goal is to study how the performance of our procedure varies when we change key parameters such as T_0, T_1, N, and the degree of serial dependence. In addition, we shed some light on the finite sample performance of our method under misspecification.

We consider different data generating processes (DGPs) for the potential outcomes of the treated unit in the absence of the treatment all of which specify the treated outcome as a weighted average of the control outcomes:

$$Y_t^N = \sum_{i=1}^{N} w_i Y_{it}^N + u_t,$$

where $u_t = \rho_u u_{t-1} + v_t$, where $v_t \overset{iid}{\sim} N(0, 1 - \rho_u^2)$. The DGPs differ with respect to the specification of the weights w.

	Weight Specification	$\|w\|_1$
DGP1	$w \propto (1, 1, 1, 0, \ldots, 0)'$	1
DGP2	$w \propto (1, \ldots, 1)'$	1
DGP3	$w \propto (1, \ldots, 1)'$	3

DGP1 represents a setting with very sparse weights. In DGP2, we set all weights equal to $w_i = 1/N$, which corresponds to the difference-in-differences model; see Remark 1. In the simulations, we choose $Q = 1$ such that DGP3 represents a setting in which constrained Lasso is misspecified.

Following Hahn and Shi (2016), we impose factor model for the control units:

$$Y_{it}^N = \mu_i + \theta_t + \lambda_i F_t + \epsilon_{it},$$

Here, we set $\mu_i = i/J$, $\lambda_i = i/J$, $\theta_t \overset{iid}{\sim} N(0,1)$, $F_t \overset{iid}{\sim} N(0,1)$, and $\epsilon_{it} = \rho_\epsilon \epsilon_{it-1} + \xi_{jt}$, where $\xi_{jt} \overset{iid}{\sim} N(0, 1-\rho_\epsilon^2)$. We consider settings with i.i.d. data ($\rho_u = \rho_\epsilon = 0$) and weakly dependent data ($\rho_u = \rho_\epsilon = 0.6$).

The treatment effects are generated as $\alpha_t = \tau_0 + \tilde{\xi}_t$, where $\tilde{\xi}_t = \xi_t - 1/T_1 \sum_{t=T_0+1}^{T} \xi_t$ and $\xi_t \overset{iid}{\sim} N(0,1)$, which ensures that the true effect is exactly equal to τ_0 in all simulated samples. We do not redraw treatments effect across simulations and set $\tau_0 = 0$.

We consider different values for K, $K \in \{2,3,4,5\}$, and compare our method to the Lasso-based ARCO approach proposed by Carvalho et al. (2017). The nominal coverage is equal to $1 - \alpha = 0.9$.

Table 2 shows the empirical coverage and the average length of the confidence intervals with i.i.d. data ($\rho_u = \rho_\epsilon = 0$). Our method exhibits very close-to-correct coverage, irrespective of the choice of K and whether or not constrained Lasso is correctly specified. As predicted by our theory, the average length of the confidence intervals is decreasing in K. In addition, we find that, under correct specification (DGP1 and DGP2), the confidence intervals are shorter than under misspecification (DGP3). By comparison, ArCo typically exhibits undercoverage and yields confidence intervals that are slightly shorter on average than the intervals obtained from our approach with $K = 5$ under correct specification.

Table 3 presents the results under weak dependence ($\rho_u = \rho_\epsilon = 0.6$). With $K = 2$, our method exhibits almost exact coverage but yields very wide confidence intervals. For $K = 3$, coverage is still very close to the nominal level, but the confidence bands are much shorter on average. As predicted by our theory, choosing $K = 4$ or $K = 5$ further reduces the average length of the confidence intervals but can lead to undercoverage, especially when $T_0 = 20$. However, we note that the coverage properties improve substantially when going from $T_0 = T_1 = 20$ to $T_0 = T_1 = 40$, suggesting that when T_0 and T_1 become larger, choosing $K = 4$ (or even $K = 5$) may be sensible in practice. Under weak dependence, ArCo yields shorter confidence intervals than our method but exhibits substantial undercoverage under all three DGP.

5 Application: the economic costs of conflict

To illustrate our approach, we analyze the economic effects of conflict. We follow Abadie and Gardeazabal (2003) and use the terrorist conflict in the Basque Country as a case study. Before the outbreak of terrorist activity by the Basque terrorist organization ETA in the early 1970's, the Basque Country was the third richest region in Spain in terms of GDP per capita. After 30 years of terrorism and political conflicts, the Basque Country had dropped to the sixth position. Our goal is to estimate the causal effect of terrorism on per-capita GDP in the Basque Country. We use the same dataset as in Abadie and Gardeazabal (2003).[2] The data contain annual real per-capita GDP at the province-level in thousands

[2]The data are available through the R-package `Synth` (Abadie et al., 2011).

of 1986 USD from 1955 to 1997. Figure 2 displays the raw data. Because the data are non-stationary, we apply our method to de-trended data.[3] There are $N + 1 = 17$ provinces in total. The terrorist activities gained in strength during the 1970's. In 1973, ETA killed Franco's Prime Minister and for the first time was responsible for more than five deaths per year; see Table 1 in Abadie and Gardeazabal (2003). Therefore, our pre-intervention period goes from 1955 to 1972 ($T_0 = 18$). Our post-intervention period goes from 1973 to 1997 ($T_1 = 25$).

Our empirical strategy uses the $N = 16$ other regions as control units. We impose the following model for the per capita GDP in the Basque Country in the absence of terrorism.

$$Y_t^N = \mu + \sum_{i=1}^{N} w_i Y_{it}^N + u_t, \quad \|w\|_1 \leq 1.$$

We estimate the ATE using our method with $K = 3$. This choice is based on the empirical Monte Carlo simulations in Section 4.1. For comparison, we also report the estimates for $K = 2$ and $K = 4$. Table 4 presents the results. For $K = 3$, the estimated effect of terrorism on GDP is negative and significant. The results for $K = 2$ and $K = 4$ are similar. We compare our results to estimates obtained from the Lasso-based ArCo approach proposed by Carvalho et al. (2017). The ArCo point estimate is somewhat smaller (in absolute value) than the estimates based on our method, but also significantly different from zero.

6 Conclusion

This paper develops new methods for making inference on average treatment effects in aggregate panel data settings. We approximate the counterfactual outcomes in the absence of the treatment using linear models estimated by constrained Lasso. The proposed inference method is based on a cross-fitting procedure for bias correction in conjunction with a self-normalized t-statistic. Our procedure is very easy to implement and has several theoretical advantages: it does not rely on sparsity, is fully robust against misspecification, avoids estimation of the long run variance, and is more efficient than difference-in-differences and difference-in-means estimators. The proposed method exhibits excellent finite sample properties in settings where T_0 and T_1 are very small. We use our method to re-evaluate the economic consequences of terrorism.

Here, we focus on the constrained Lasso estimator because it nests popular existing approaches, is essentially tuning free, does not rely on sparsity, has desirable theoretical properties, and admits a characterization under misspecification. However, many of the ideas presented in this paper such as the cross-fitting scheme and the construction of the test statistic are generic and could be used in conjunction with other CSC methods for estimating counterfactuals.

[3] We subtract a linear trend which is estimated based on the pre-treatment period for the treated unit and based on all periods for the control units.

References

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of californias tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Abadie, A., Diamond, A., and Hainmueller, J. (2011). Synth: An R package for synthetic control methods in comparative case studies. *Journal of Statistical Software*, 42(13):1–17.

Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510.

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *The American Economic Review*, 93(1):113–132.

Amjad, M. J., Shah, D., and Shen, D. (2017). Robust synthetic control.

Angrist, J. and Pischke, S. (2008). *Mostly Harmless Econometrics: An Empiricists' Companion*. Princeton University Press.

Ashenfelter, O. and Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics*, 67(4):648–660.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2017). Matrix completion methods for causal panel data models.

Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497.

Athreya, K. B. and Lahiri, S. N. (2006). *Measure Theory and Probability Theory*. Springer Science & Business Media.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates?*. *The Quarterly Journal of Economics*, 119(1):249–275.

Bloniarz, A., Liu, H., Zhang, C.-H., Sekhon, J. S., and Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27):7383–7390.

Card, D. and Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *The American Economic Review*, 84(4):772–793.

Carvalho, C. V., Masini, R., and Medeiros, M. C. (2017). Arco: an artificial counterfactual approach for high-dimensional panel time-series data.

Chan, M. and Kwok, S. (2016). Policy evaluation with interactive fixed effects.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2017). An exact and robust conformal inference method for counterfactual and synthetic controls. arXiv:1712.09089.

Doudchenko, N. and Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Working Paper 22791, National Bureau of Economic Research.

Ferman, B. and Pinto, C. (2017). Placebo tests for synthetic controls.

Firpo, S. and Possebom, V. (2017). Synthetic control method: Inference, sensitivity analysis and confidence sets.

Gobillon, L. and Magnac, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *The Review of Economics and Statistics*, 98(3):535–551.

Hahn, J. and Shi, R. (2016). Synthetic control and inference. Mimeo.

Hsiao, C., Steve Ching, H., and Ki Wan, S. (2012). A panel data approach for program evaluation: Measuring the benefits of political and economic integration of hong kong with mainland china. *Journal of Applied Econometrics*, 27(5):705–740.

Ibragimov, R. and Müller, U. K. (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics*, 28(4):453–468.

Kim, D. and Oka, T. (2014). Divorce law reforms and divorce rates in the usa: An interactive fixed-effects approach. *Journal of Applied Econometrics*, 29(2):231–245.

Lei, L. and Ding, P. (2018). Regression adjustment in randomized experiments with a diverging number of covariates. *arXiv:1806.07585*.

Li, K. (2018). Inference for factor model based average treatment effects.

Li, K. T. (2017). Statistical inference for average treatment effects estimated by synthetic control methods.

Li, K. T. and Bell, D. R. (2017). Estimation of average treatment effects with panel data: Asymptotic theory and implementation. *Journal of Econometrics*, 197(1):65 – 75.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. *Statistical Science*, Reprint, 5:463–480.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012.

Raskutti, G., Wainwright, M. J., and Yu, B. (2010). Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(Aug):2241–2259.

Raskutti, G., Wainwright, M. J., and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on Information Theory*, 57(10):6976–6994.

Rubin, D. B. (1974). Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447.

Valero, R. (2015). Synthetic control method versus standard statistic techniques a comparison for labor market reforms.

White, H. (2014). *Asymptotic theory for econometricians*. Academic press.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76.

A Proofs

A.1 Proof of Theorem 1

We use the notation $E_{T,(-k)}$ for $|H_{(-k)}|^{-1}\sum_{t\in H_{(-k)}}$. To simplify notations, we use $\delta = \hat{w}_{(k)} - w$ instead of $\Delta_{(k)} = \hat{w}_{(k)} - w$. Let $\eta_1 = \|E_{T,(-k)}(X_tX_t' - E(X_tX_t'))\|_\infty$ and $\eta_2 = \|E_{T,(-k)}X_tu_t\|_\infty$.

Notice that $E_{T,(-k)}(u_t - X_t'\delta)^2 \le E_{T,(-k)}u_t^2$. Therefore,

$$E_{T,(-k)}(X_t'\delta)^2 \le 2(E_{T,(-k)}X_tu_t)'\delta \le 2\eta_2\|\delta\|_1.$$

Notice that

$$\left|\delta'\left(E_{T,(-k)}X_tX_t' - \Sigma_{(-k)}\right)\delta\right| \le \|\delta\|_1^2\left\|E_{T,(-k)}X_tX_t' - \Sigma_{(-k)}\right\|_\infty = \eta_1\|\delta\|_1^2.$$

This means that

$$E_{T,(-k)}(X_t'\delta)^2 \ge \delta'\Sigma_{(-k)}\delta - \eta_1\|\delta\|_1^2 \ge \kappa\|\delta\|_2^2 - \eta_1\|\delta\|_1^2.$$

It follows that

$$\kappa\|\delta\|_2^2 - \eta_1\|\delta\|_1^2 \le 2\eta_2\|\delta\|_1.$$

Since $\|\delta\|_1 \le \|\hat{w}_{(k)}\|_1 + \|w\|_1 \le 2Q$, we obtain the result by $\|\delta\|_2^2 \le (\eta_1\|\delta\|_1^2 + 2\eta_2\|\delta\|_1)/\kappa$. □

A.2 Proof of Lemma 1

Let $\mu = E(X_t)$. Notice that for $T_0 + 1 \le t \le T$,

$$Y_t - X_t'\hat{w}_{(k)} = \alpha_t + u_t - X_t'\Delta_{(k)} = \alpha_t + u_t - \mu'\Delta_{(k)} - \tilde{X}_t'\Delta_{(k)}$$

and for $t \in H_k$

$$Y_t - X_t'\hat{w}_{(k)} = u_t - X_t'\Delta_{(k)} = u_t - \mu'\Delta_{(k)} - \tilde{X}_t'\Delta_{(k)}.$$

Therefore,

$$\begin{aligned}
\hat{\tau}_k - \tau &= T_1^{-1}\sum_{t=T_0+1}^{T}(Y_t - X_t'\hat{w}_{(k)}) - \frac{1}{|H_k|}\sum_{t\in H_k}(Y_t - X_t'\hat{w}_{(k)}) - \tau \\
&= T_1^{-1}\sum_{t=T_0+1}^{T}(\alpha_t + u_t - \mu'\Delta_{(k)} - \tilde{X}_t'\Delta_{(k)}) - \frac{1}{|H_k|}\sum_{t\in H_k}(u_t - \mu'\Delta_{(k)} - \tilde{X}_t'\Delta_{(k)}) - \tau \\
&= T_1^{-1}\sum_{t=T_0+1}^{T}u_t - \frac{1}{|H_k|}\sum_{t\in H_k}u_t + \frac{1}{|H_k|}\sum_{t\in H_k}\tilde{X}_t'\Delta_{(k)} - T_1^{-1}\sum_{t=T_0+1}^{T}\tilde{X}_t'\Delta_{(k)}.
\end{aligned}$$

The proof is complete. □

A.3 Proof of Theorem 2

Fix $k \in \{1, ..., K\}$. Define B_k to be the "two-sided buffer", i.e., the set that contains the smallest γ_T numbers and the largest γ_T numbers in H_k. Also define $A_k = H_k \backslash B_k$, i.e., $A_k = \{t : \min H_k + \gamma_T + 1 \leq t \leq \max H_k - \gamma_T\}$. Thus,

$$\sum_{t \in H_k} \tilde{X}_t' \Delta_{(k)} = \sum_{t \in A_k} \tilde{X}_t' \Delta_{(k)} + \sum_{t \in B_k} \tilde{X}_t' \Delta_{(k)}.$$

The second term can be bounded by

$$\left| \sum_{t \in B_k} \tilde{X}_t' \Delta_{(k)} \right| \leq \max_{1 \leq t \leq T_0} \|\tilde{X}_t\|_\infty \|\Delta_{(k)}\|_1 |B_k| = 2\gamma_T \max_{1 \leq t \leq T_0} \|\tilde{X}_t\|_\infty \|\Delta_{(k)}\|_1.$$

Thus,

$$P\left(\left| \sum_{t \in B_k} \tilde{X}_t' \Delta_{(k)} \right| \leq 2\rho_T \gamma_T \|\Delta_{(k)}\|_1 \right) \to 1. \tag{13}$$

On the other hand, we use Berbee's coupling to bound $\sum_{t \in A_k} \tilde{X}_t' \Delta_{(k)}$. By Theorem 16.2.1 of Athreya and Lahiri (2006), on an enlarged probability space, there exist random variables $\{\bar{X}_t\}_{t \in A_k}$ such that (1) $\{\bar{X}_t\}_{t \in A_k}$ and $\{\tilde{X}_t\}_{t \in A_k}$ have the same distribution, (2) $\{\bar{X}_t\}_{t \in A_k}$ is independent of data in $\{1, ..., T_0\} \backslash H_k$ and (3) $P(\{\bar{X}_t\}_{t \in A_k} \neq \{\tilde{X}_t\}_{t \in A_k}) \leq \beta_{\text{mix}}(\gamma_T)$. Notice that $\Delta_{(k)}$ is independent of $\{\bar{X}_t\}_{t \in A_k}$. Hence,

$$E\left[\left(\sum_{t \in A_k} \bar{X}_t' \Delta_{(k)} \right)^2 \mid \Delta_{(k)} \right] = \Delta_{(k)}' E\left[\left(\sum_{t \in A_k} \bar{X}_t \right) \left(\sum_{t \in A_k} \bar{X}_t \right)' \right] \Delta_{(k)} \overset{\text{(i)}}{\leq} |A_k| \kappa_1 \|\Delta_{(k)}\|_2^2,$$

where (i) follows by Assumption 4 and the fact that $\{\bar{X}_t\}_{t \in A_k}$ and $\{\tilde{X}_t\}_{t \in A_k}$ have the same distribution. Thus, $\sum_{t \in A_k} \bar{X}_t' \Delta_{(k)} = O_P(\sqrt{|A_k|} \|\Delta_{(k)}\|_2)$. Since $P(\{\bar{X}_t\}_{t \in A_k} \neq \{\tilde{X}_t\}_{t \in A_k}) \leq \beta_{\text{mix}}(\gamma_T) = o(1)$, it follows that

$$\sum_{t \in A_k} \tilde{X}_t' \Delta_{(k)} = O_P(\sqrt{|A_k|} \|\Delta_{(k)}\|_2). \tag{14}$$

Now by (13) and (14),

$$\sum_{t \in H_k} \tilde{X}_t' \Delta_{(k)} = O_P\left(\rho_T \gamma_T \|\Delta_{(k)}\|_1 + \sqrt{|A_k|} \|\Delta_{(k)}\|_2 \right) \overset{\text{(i)}}{=} O_P\left(\rho_T \gamma_T \max_{1 \leq k \leq K} \|\Delta_{(k)}\|_1 + \sqrt{T_0} \max_{1 \leq k \leq K} \|\Delta_{(k)}\|_2 \right),$$

where (i) follows by the assumption that $\gamma_T = o(T_0)$, $\min_{1 \leq k \leq K} |H_k|/T_0 \geq 1/(K+1)$ and K is bounded. Thus,

$$T_0^{-1} \sum_{k=1}^{k} \sum_{t \in H_k} \tilde{X}_t' \Delta_{(k)} = O_P\left(T_0^{-1} \rho_T \gamma_T \max_{1 \leq k \leq K} \|\Delta_{(k)}\|_1 + T_0^{-1/2} \max_{1 \leq k \leq K} \|\Delta_{(k)}\|_2 \right).$$

Similarly, we can show

$$T_1^{-1}\sum_{t=T_0+1}^{T}\tilde{X}_t'\Delta = O_P\left(T_1^{-1}\rho_T\gamma_T\max_{1\le k\le K}\|\Delta_{(k)}\|_1 + T_1^{-1/2}\max_{1\le k\le K}\|\Delta_{(k)}\|_2\right).$$

The desired result follows by the assumption $\rho_T\gamma_T = o(\min\{\sqrt{T_0},\sqrt{T_1}\})$ and the fact that $\max_{1\le k\le K}\|\Delta_{(k)}\|_1$ is bounded (due to the assumption of $\|w\|_1 = O(1)$). □

A.4 Proof of Theorem 3

Part (i) is a direct consequence of Corollary 1. For Part (ii), Corollary 1 and the continuous mapping theorem imply that

$$\mathbb{T}_K \xrightarrow{d} \mathcal{T}_K,$$

where

$$\begin{aligned}
\mathcal{T}_K &= \frac{\sqrt{K}(K^{-1}\sum_{k=1}^K(\sqrt{c_0}\xi_0 - \sqrt{K}\xi_k))}{\sqrt{1+c_0}\sqrt{(K-1)^{-1}\sum_{k=1}^K((\sqrt{c_0}\xi_0 - \sqrt{K}\xi_k) - K^{-1}\sum_{k=1}^K(\sqrt{c_0}\xi_0 - \sqrt{K}\xi_k))^2}} \\
&= \frac{(1+c_0)^{-1/2}(\sqrt{Kc_0}\xi_0 - K\bar{\xi})}{\sqrt{(K-1)^{-1}\sum_{k=1}^K(\sqrt{K}(\xi_k - \bar{\xi}))^2}} \\
&= \frac{(1+c_0)^{-1/2}(\sqrt{c_0}\xi_0 - \sqrt{K}\bar{\xi})}{\sqrt{(K-1)^{-1}\sum_{k=1}^K(\xi_k - \bar{\xi})^2}},
\end{aligned}$$

where $\bar{\xi} = K^{-1}\sum_{k=1}^K\xi_k$. Notice that $\sum_{k=1}^K(\xi_k-\bar{\xi})^2$ is independent of $\bar{\xi}$ and thus is independent of the numerator of $\mathcal{T}_K$. It follows that $\mathcal{T}_K$ has a student t-distribution with $K-1$ degrees of freedom. □

A.5 Proof of Theorem 4

For simplicity, we write $\hat{w} = \hat{w}_{(k)}$, $\hat{\mu} = \hat{\mu}_{(-k)}$ and $\hat{\Sigma} = \hat{\Sigma}_{(-k)}$. Let $\xi = (\hat{\mu} - \hat{\Sigma}w_*) - (\mu - \Sigma w_*)$. By assumption, $\|\xi\|_\infty = o_P(1)$.

We rewrite

$$w_* = \arg\min_v \; v'\Sigma v - 2\mu' v \qquad s.t. \qquad \|v\|_1 \le Q$$

and

$$\hat{w} = \arg\min_v \; v'\hat{\Sigma} v - 2\hat{\mu}' v \qquad s.t. \qquad \|v\|_1 \le Q.$$

Let $\Delta = \hat{w} - w_*$. For any $\lambda \in [0,1]$, define $w_\lambda = w_* + \lambda\Delta$. Then by the definition of w_*, we have that $w_\lambda'\Sigma w_\lambda - 2\mu' w_\lambda \ge w_*'\Sigma w_* - 2\mu' w_*$, which means

$$\lambda^2\Delta'\Sigma\Delta \ge 2\lambda(\mu - \Sigma w_*)'\Delta.$$

Thus, for any $\lambda \in (0,1)$, we have that $\lambda\Delta'\Sigma\Delta \geq 2(\mu - \Sigma w_*)'\Delta$. Since this holds for any $\lambda \in (0,1)$, we have that

$$(\mu - \Sigma w_*)'\Delta \leq 0. \tag{15}$$

Now by definition, we have that

$$\hat{\beta}'\hat{\Sigma}\hat{\beta} - 2\hat{\mu}'\hat{\beta} \leq w_*'\hat{\Sigma}w_* - 2\hat{\mu}'w_*.$$

It follows that

$$\Delta'\hat{\Sigma}\Delta \leq 2(\hat{\mu} - \hat{\Sigma}w_*)'\Delta.$$

Notice that

$$(\hat{\mu} - \hat{\Sigma}w_*)'\Delta = (\mu - \Sigma w_*)'\Delta + \xi'\Delta \overset{\text{(i)}}{\leq} \xi'\Delta \leq \|\xi\|_\infty \|\Delta\|_1 \leq 2\|\xi\|_\infty Q,$$

where (i) holds by (15). The above two displays imply that

$$\Delta'\hat{\Sigma}\Delta \leq 2\|\xi\|_\infty Q.$$

Therefore

$$2\|\xi\|_\infty Q \geq \Delta'\Sigma\Delta + \Delta'(\hat{\Sigma} - \Sigma)\Delta \geq \Delta'\Sigma\Delta - \|\hat{\Sigma} - \Sigma\|_\infty \|\Delta\|_1^2 \geq c\|\Delta\|_2^2 - 4\|\hat{\Sigma} - \Sigma\|_\infty Q^2.$$

The desired result follows. □

A.6 Proof of Theorem 5

In view of Corollary 2, the proof follows by the same arguments as in the proof of Theorem 3. □

A.7 Proof of Theorem 6

It is easy to see that

$$\tilde{\tau} - \tau = T_1^{-1} \sum_{t=T_0+1}^{T} Y_t^N - T_0^{-1} \sum_{t=1}^{T_0} Y_t^N.$$

By Lyapunov's central limit theorem, we have that $\sqrt{T_0}(\tilde{\tau} - \tau) \xrightarrow{d} N(0, (c_0+1)\sigma_0^2)$. For the second claim, notice that in this case

$$\sigma_*^2 = E(Y_t^N - X_t'w_*)^2 = \min_{\|v\|_1 \leq Q} E(Y_t^N - X_t'v)^2 \leq E(Y_t^N)^2 = \sigma_0^2.$$

The proof is complete. □

B Tables and figures

Table 1: Empirical Monte Carlo simulations

Coverage				
$K = 2$	$K = 3$	$K = 4$	$K = 5$	ArCo
0.91	0.84	0.74	0.67	0.57
Average length				
0.62	0.22	0.15	0.12	0.11

Notes: Simulation design based on the empirical application as described in the main text. Nominal coverage $1 - \alpha = 0.9$. Based on simulations with 1000 repetitions.

Table 2: Results i.i.d. data

Coverage

$J = 20$

		DGP1					DGP2					DGP3				
T_0	T_1	$K=2$	$K=3$	$K=4$	$K=5$	ArCo	$K=2$	$K=3$	$K=4$	$K=5$	ArCo	$K=2$	$K=3$	$K=4$	$K=5$	ArCo
20	20	0.92	0.92	0.92	0.90	0.85	0.89	0.92	0.90	0.90	0.84	0.89	0.90	0.90	0.88	0.78
40	20	0.91	0.92	0.90	0.89	0.87	0.92	0.91	0.91	0.91	0.84	0.91	0.91	0.91	0.91	0.85
20	40	0.91	0.90	0.91	0.89	0.88	0.90	0.91	0.89	0.89	0.86	0.90	0.89	0.88	0.89	0.80
40	40	0.90	0.92	0.90	0.90	0.93	0.91	0.90	0.89	0.91	0.93	0.90	0.90	0.90	0.91	0.89

$J = 40$

		DGP1					DGP2					DGP3				
T_0	T_1	$K=2$	$K=3$	$K=4$	$K=5$	ArCo	$K=2$	$K=3$	$K=4$	$K=5$	ArCo	$K=2$	$K=3$	$K=4$	$K=5$	ArCo
20	20	0.90	0.91	0.88	0.88	0.81	0.92	0.92	0.90	0.90	0.85	0.89	0.90	0.90	0.88	0.74
40	20	0.91	0.91	0.91	0.90	0.85	0.93	0.92	0.89	0.91	0.86	0.90	0.90	0.91	0.89	0.84
20	40	0.90	0.90	0.91	0.90	0.85	0.91	0.92	0.91	0.92	0.88	0.90	0.90	0.90	0.91	0.81
40	40	0.91	0.90	0.90	0.92	0.93	0.90	0.92	0.90	0.91	0.93	0.89	0.90	0.90	0.90	0.89

Average length

$J = 20$

		DGP1					DGP2					DGP3				
T_0	T_1	$K=2$	$K=3$	$K=4$	$K=5$	ArCo	$K=2$	$K=3$	$K=4$	$K=5$	ArCo	$K=2$	$K=3$	$K=4$	$K=5$	ArCo
20	20	4.18	2.17	1.67	1.51	1.26	4.02	2.07	1.61	1.44	1.20	8.46	4.55	3.52	3.21	1.27
40	20	3.50	1.63	1.31	1.21	1.09	3.44	1.65	1.32	1.23	1.06	7.20	3.63	3.03	2.76	1.12
20	40	3.53	1.83	1.42	1.30	1.16	3.53	1.79	1.38	1.25	1.15	6.86	3.84	3.00	2.84	1.14
40	40	2.75	1.37	1.10	1.02	1.02	2.68	1.31	1.07	0.98	1.00	5.72	2.98	2.43	2.29	1.01

$J = 40$

		DGP1					DGP2					DGP3				
T_0	T_1	$K=2$	$K=3$	$K=4$	$K=5$	ArCo	$K=2$	$K=3$	$K=4$	$K=5$	ArCo	$K=2$	$K=3$	$K=4$	$K=5$	ArCo
20	20	4.15	2.22	1.71	1.57	1.27	4.02	2.10	1.58	1.45	1.21	8.30	4.26	3.36	3.15	1.22
40	20	3.65	1.77	1.41	1.27	1.09	3.52	1.67	1.32	1.23	1.06	6.98	3.57	2.95	2.70	1.07
20	40	3.60	1.92	1.45	1.33	1.18	3.33	1.79	1.36	1.24	1.10	6.90	3.70	2.89	2.76	1.11
40	40	2.76	1.36	1.11	1.03	1.03	2.59	1.33	1.08	0.99	0.99	5.75	2.82	2.42	2.24	1.01

Notes: Simulation design as described in the main text with $\rho_u = \rho_\epsilon = 0$. Nominal coverage $1 - \alpha = 0.9$. Based on simulations with 1000 repetitions.

Table 3: Results weakly dependent data

Coverage

$J = 20$

		DGP1					DGP2					DGP3				
T_0	T_1	$K=2$	$K=3$	$K=4$	$K=5$	ArCo	$K=2$	$K=3$	$K=4$	$K=5$	ArCo	$K=2$	$K=3$	$K=4$	$K=5$	ArCo
20	20	0.90	0.88	0.83	0.79	0.73	0.90	0.88	0.84	0.81	0.72	0.92	0.90	0.85	0.87	0.65
40	20	0.91	0.88	0.87	0.87	0.76	0.92	0.90	0.88	0.86	0.74	0.91	0.90	0.90	0.89	0.71
20	40	0.90	0.87	0.83	0.82	0.76	0.90	0.89	0.84	0.81	0.76	0.90	0.90	0.86	0.86	0.68
40	40	0.90	0.90	0.87	0.84	0.82	0.89	0.88	0.88	0.86	0.83	0.89	0.90	0.90	0.89	0.78

$J = 40$

		DGP1					DGP2					DGP3				
T_0	T_1	$K=2$	$K=3$	$K=4$	$K=5$	ArCo	$K=2$	$K=3$	$K=4$	$K=5$	ArCo	$K=2$	$K=3$	$K=4$	$K=5$	ArCo
20	20	0.91	0.87	0.84	0.80	0.72	0.90	0.89	0.85	0.83	0.73	0.90	0.89	0.87	0.86	0.68
40	20	0.90	0.89	0.88	0.85	0.74	0.91	0.90	0.89	0.86	0.75	0.90	0.89	0.88	0.87	0.71
20	40	0.90	0.86	0.83	0.81	0.75	0.90	0.87	0.84	0.80	0.75	0.88	0.89	0.88	0.87	0.71
40	40	0.89	0.88	0.86	0.83	0.79	0.90	0.88	0.87	0.85	0.80	0.91	0.90	0.89	0.90	0.73

Average length

$J = 20$

		DGP1					DGP2					DGP3				
T_0	T_1	$K=2$	$K=3$	$K=4$	$K=5$	ArCo	$K=2$	$K=3$	$K=4$	$K=5$	ArCo	$K=2$	$K=3$	$K=4$	$K=5$	ArCo
20	20	6.55	3.08	2.22	1.90	3.40	6.54	2.97	2.20	1.88	1.55	10.47	5.01	3.91	3.53	1.76
40	20	6.20	2.89	2.27	2.01	1.63	6.29	2.87	2.28	1.98	1.46	9.08	4.50	3.62	3.32	1.58
20	40	5.71	2.71	1.97	1.67	1.46	5.23	2.50	1.82	1.59	1.36	8.69	4.38	3.37	2.99	1.46
40	40	4.83	2.41	1.88	1.63	1.37	4.81	2.29	1.81	1.58	1.31	7.37	3.74	2.96	2.73	1.37

$J = 40$

		DGP1					DGP2					DGP3				
T_0	T_1	$K=2$	$K=3$	$K=4$	$K=5$	ArCo	$K=2$	$K=3$	$K=4$	$K=5$	ArCo	$K=2$	$K=3$	$K=4$	$K=5$	ArCo
20	20	6.83	3.16	2.29	1.98	1.72	6.27	2.94	2.09	1.84	1.54	10.42	5.00	3.88	3.50	1.68
40	20	6.13	2.96	2.31	2.00	1.52	6.28	2.86	2.20	1.93	1.45	9.05	4.41	3.54	3.19	1.51
20	40	5.62	2.64	1.98	1.70	1.49	5.30	2.56	1.85	1.59	1.35	8.56	4.43	3.36	3.01	1.44
40	40	4.88	2.34	1.83	1.61	1.34	4.81	2.29	1.77	1.57	1.27	7.24	3.56	2.88	2.66	1.32

Notes: Simulation design as described in the main text with $\rho_u = \rho_\epsilon = 0.6$. Nominal coverage $1 - \alpha = 0.9$. Based on simulations with 1000 repetitions.

Table 4: Results application

Method	ATE	90%-CI	
$K = 2$	-0.90	-1.22	-0.57
$K = 3$	-0.92	-1.13	-0.70
$K = 4$	-0.88	-1.00	-0.75
ArCo	-0.76	-1.16	-0.36

Figure 1: Expected length confidence interval

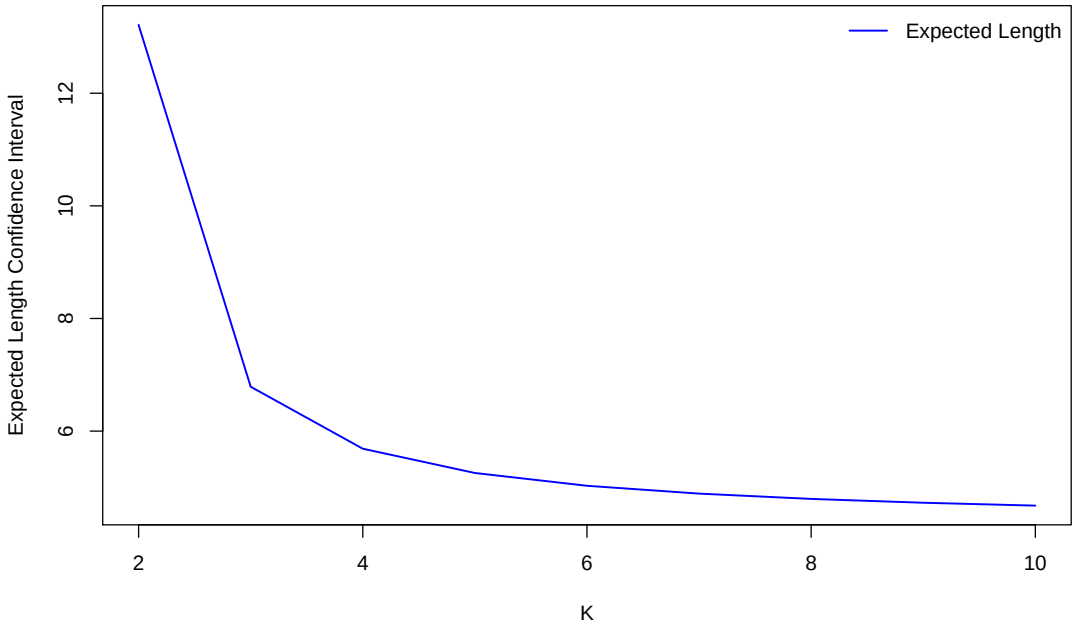

Figure 2: Raw Data