




An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls

Victor Chernozhukov, Kaspar Wüthrich & Yinchu Zhu

To cite this article: Victor Chernozhukov, Kaspar Wüthrich & Yinchu Zhu (2021): An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls, Journal of the American Statistical Association, DOI: [10.1080/01621459.2021.1920957](https://doi.org/10.1080/01621459.2021.1920957)

To link to this article: <https://doi.org/10.1080/01621459.2021.1920957>

 View supplementary material [↗](#)

 Accepted author version posted online: 23 Apr 2021.

 Submit your article to this journal [↗](#)

 Article views: 6

 View related articles [↗](#)

 View Crossmark data [↗](#)

An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls

Victor Chernozhukov*, Kaspar Wüthrich†, Yinchu Zhu‡

*Massachusetts Institute of Technology, 50 Memorial Drive, E52-361B,
Cambridge, MA 02142, USA;

†Department of Economics, University of California San Diego, 9500 Gilman Dr.,
La Jolla, CA 92093, USA;

‡Brandeis University, 415 South Street, Waltham, MA 02453, USA;

Email: kwuthrich@ucsd.edu

Abstract

We introduce new inference procedures for counterfactual and synthetic control methods for policy evaluation. We recast the causal inference problem as a counterfactual prediction and a structural breaks testing problem. This allows us to exploit insights from conformal prediction and structural breaks testing to develop permutation inference procedures that accommodate modern high-dimensional estimators, are valid under weak and easy-to-verify conditions, and are provably robust against misspecification. Our methods work in conjunction with many different approaches for predicting counterfactual mean outcomes in the absence of the policy intervention. Examples include synthetic controls, difference-in-differences, factor and matrix completion models, and (fused) time series panel data models. Our approach demonstrates an excellent small-sample performance in simulations and is taken to a data application where we re-evaluate the consequences of decriminalizing indoor prostitution. Open-source software for implementing our conformal inference methods is available.

Keywords: permutation inference, model-free validity, difference-in-differences, factor model, matrix completion, constrained Lasso

1 Introduction

We consider the problem of making inferences on the causal effect of a policy intervention in an aggregate time series setup with a single treated unit. The

treated unit is observed for T_0 periods before and T_* periods after the intervention occurs. Often, there is additional information in the form of possibly very many untreated units, which can serve as controls. Such settings are ubiquitous in applied research, and there are many different approaches for estimating the causal effect of the policy. Examples include difference-in-differences methods, synthetic control (SC) approaches, factor, matrix completion, and interactive fixed effects (FE) models, and times series models.¹ We refer to these methods as counterfactual and synthetic control (CSC) methods.

This paper provides generic and robust procedures for making inferences on policy effects estimated by CSC methods. We propose a general counterfactual modeling framework that nests and generalizes many traditional and new methods for counterfactual analysis. We focus on methods that are able to generate mean-unbiased proxies, P_t^N , for the counterfactual outcomes of the treated unit in the absence of the policy intervention, Y_{lt}^N :

$$Y_{lt}^N = P_t^N + u_t, \quad E(u_t) = 0, \quad t = 1, \dots, T_0 + T_*.$$

The policy effect in period t is $\theta_t = Y_{lt}^I - Y_{lt}^N$, where Y_{lt}^I is the counterfactual outcome of the treated unit with the policy intervention. We are interested in testing hypotheses about the trajectory of policy effects in the post-treatment period, $\theta = \{\theta_t\}_{t=T_0+1}^{T_0+T_*}$. Specifically, we postulate a trajectory $\theta^0 = \{\theta_t^0\}_{t=T_0+1}^{T_0+T_*}$ and test the sharp null hypothesis that $\theta = \theta^0$. We also consider the problem of testing hypotheses about per-period effects θ_t and propose a simple algorithm for constructing pointwise confidence intervals via test inversion.

We recast the inference problem as a (counterfactual) prediction and a structural breaks testing problem. This allows us to build on the literature on conformal prediction (Vovk et al., 2005) and end-of-sample stability testing (Dufour et al., 1994; Andrews, 2003) to construct inference procedures that are provably robust against misspecification and accommodate many classical and modern high-dimensional methods for estimating P_t^N .

The basic idea of our testing procedures is as follows. Suppose that there is only one post-treatment period and that P_t^N is known. Under the sharp null that $\theta_{T_0+1} = \theta_{T_0+1}^0$, we can compute Y_{1t}^N and $u_t = Y_{1t}^N - P_t^N$ for all time periods. If the stochastic shock process $\{u_t\}$ is stationary and weakly dependent, and its distribution is invariant under the intervention, the distribution of the error in the post-treatment period, u_{T_0+1} , should be the same as the distribution of the errors in the pre-treatment period, $\{u_t\}_{t=1}^{T_0}$. We operationalize this idea by proposing inference methods in which p -values are obtained by permuting blocks of estimated residuals across the time series dimension.

The proposed methods are valid under two different sets of conditions:

(i) Estimator Consistency and Stationary Weakly Dependent Errors

If the data exhibit dynamics, trends, and serial dependence but the stochastic shock sequence $\{u_t\}$ is stationary and weakly dependent, our inference procedures are approximately valid if the estimator of P_t^N is consistent (pointwise and in prediction norm). Consistency can be verified for many different CSC methods. We provide concrete sufficient conditions for a representative selection of methods, including difference-in-differences, SC, factor models, matrix completion and interactive FE models, linear and nonlinear time series models, and fused time series panel data models.

(ii) Estimator Stability and Stationary Weakly Dependent Data

In practice, misspecification is an important concern. We show that even if the model for P_t^N is misspecified and the estimator of P_t^N , \hat{P}_t^N , is inconsistent, our procedures are still valid, provided that the data are stationary and weakly dependent and \hat{P}_t^N satisfies a *stability* condition. This condition requires that \hat{P}_t^N is stable under perturbations in a few observations. It is implied, for instance, if \hat{P}_t^N is consistent for a pseudo-

true parameter value but is shown to hold even in high-dimensional settings where consistency results under misspecification are often not available.

The main theoretical results in this paper are finite sample (non-asymptotic) bounds on the size accuracy of our methods; these bounds imply that our methods are exact as $T_0 \rightarrow \infty$. Unlike traditional asymptotic results, which are only informative when the sample size is large enough, our non-asymptotic bounds show how different factors affect the finite sample performance. This feature is relevant in CSC applications where sample sizes are often small.

A key feature of our conformal inference methods is that P_t^N is estimated under the null hypothesis based on data from all $T_0 + T_*$ periods. Estimation under the null guarantees the exact finite sample validity of our procedures if the data are iid or exchangeable. Even when exchangeability fails, imposing the null for estimation is essential for a good performance in CSC applications where T_0 is often rather small. Figure 1 plots the empirical rejection probabilities for testing the null that $\theta_{T_0+1} = 0$ when $T_0 = 19$, $J = 50$ (as in our empirical application), $\{u_t\}$ is an AR(1) process, and P_t^N is estimated using SC. The size properties of our method are excellent. By contrast, estimating P_t^N based on the T_0 pre-treatment periods without imposing the null yields substantial size distortions. Figure 1 suggests that imposing the null continues to improve size accuracy even when exchangeability fails and that these improvements can be substantial in small samples.

[Figure 1 around here.]

We make two additional contributions that may be of independent interest. First, we introduce the ℓ_1 -constrained least squares estimator or *constrained Lasso* (e.g., [Raskutti et al., 2011](#)) as an essentially tuning-free alternative to existing penalized regression estimators and study its theoretical properties. Constrained Lasso nests SC and difference-in-differences, providing a unifying approach for

the regression-based estimation of the mean proxies P_t^N . Second, we obtain theoretical consistency results for SC estimators in settings with potentially very many control units.

We develop three extensions of our main results. First, we show that our method can be modified to test hypotheses about average effects over time. Second, we extend our method to settings with multiple treated units. Third, we propose easy-to-implement placebo tests for assessing the credibility of inferences based on our method.

Monte Carlo simulations suggest that our procedures exhibit excellent size properties and are robust to misspecification. We find that imposing additional constraints (e.g., using SC instead of the more general constrained Lasso) does not improve power when these additional restrictions are correct but can cause power losses when they are not.

Finally, we re-analyze the causal effect of decriminalizing indoor prostitution on sexually transmitted infections. Following [Cunningham and Shah \(2018\)](#), we exploit the unanticipated decriminalization of indoor prostitution in Rhode Island in 2003. We find that decriminalizing indoor prostitution significantly decreased the incidence of female gonorrhea.

Related Literature.

We contribute to the literature on inference procedures for CSC methods with few treated units. A popular method is the finite population permutation approach of [Abadie et al. \(2010\)](#), see also [Firpo and Possebom \(2018\)](#) and [Abadie \(2019\)](#). This approach permutes the policy assignment and relies on permutation distributions for inference. It corresponds to conventional randomization inference ([Fisher, 1935](#)) under random assignment of the policy (e.g., [Abadie et al., 2010](#); [Abadie, 2019](#)). However, random assignment is not plausible in typical CSC applications, and assignment mechanisms are difficult to model and estimate when there are only few treated units (e.g., [Abadie, 2019](#)). [Shaikh and](#)

Toulis (2019) propose randomization tests for settings with staggered treatment adoption, which encompass the approach of Abadie et al. (2010). The main assumption of Shaikh and Toulis (2019)'s approach is that policy adoption follows a Cox proportional hazards model. We do not model the assignment mechanism. Instead, we exploit stationarity and weak dependence of the errors across time in a repeated sampling framework. One advantage of exploiting the time series dimension is that we only require a suitable model for the potential outcome of the treated unit. By contrast, cross-sectional approaches often require estimating models for all units. That is, we only require a good "local" instead of a good "global" fit, which reduces the risk of model misspecification. On the other hand, our approach requires a large number of pre-treatment periods and relies on invariance of the error distribution under the intervention.

There is also an active literature on asymptotic inference methods for CSC models. Several papers focus on testing hypotheses about average or expected effects over time, requiring T_0 and T_* to be large. Li and Bell (2017), Carvalho et al. (2018), Chernozhukov et al. (2019), and Li (2020) introduce inference methods based on penalized and constrained regression methods. Arkhangelsky et al. (2018) propose inference methods for a version of SC with time and unit weights, which admits a weighted regression formulation. Asymptotic inference methods based on factor and interactive FE models are proposed by Hsiao et al. (2012), Gobillon and Magnac (2016), Chan and Kwok (2016), Li and Bell (2017), Xu (2017), and Li (2018). Here we focus on sharp null hypotheses and permutation distributions and provide non-asymptotic performance guarantees. Our approach is generic and valid with many different methods, including constrained regression, factor models, and interactive FE estimators. Conley and Taber (2011) propose inference methods for difference-in-difference settings with few treated units. They exploit the cross-sectional dimension, relying on weak dependence and stationarity of the error terms across units, which may be hard to justify in typical CSC settings. By contrast, our procedures rely on stationarity and weak dependence of the errors over time. On the other

hand, exploiting the time series dimension, our approach requires T_0 to be large, whereas [Conley and Taber \(2011\)](#) allow T_0 to be fixed. In related work, [Cattaneo et al. \(2021\)](#) provide prediction intervals for per-period effects estimated by SC methods. Their key observation is that there is both randomness from estimating the SC weights and from the prediction error. They propose a sampling-based inference method based on non-asymptotic probability bounds that accounts for both types of randomness and is valid with stationary and non-stationary data.

We recast the causal inference problem as a (counterfactual) prediction problem and build on the literature on conformal prediction (e.g., [Vovk et al., 2005, 2009](#); [Lei et al., 2013](#); [Lei and Wasserman, 2014](#); [Lei et al., 2018](#)) and on the literature on permutation tests (e.g., [Romano, 1990](#); [Lehmann and Romano, 2005](#)), which was started by [Fisher \(1935\)](#) in the context of randomization; see [Rubin \(1984\)](#) for a Bayesian justification. On a more general level, our approach is also connected to transformation-based approaches to model-free prediction (e.g., [Politis, 2015](#)). Let us discuss in more detail the relationship to ([Chernozhukov et al., 2018](#), CWZ18 henceforth), who extend classical conformal prediction to time series settings. Besides a different focus (prediction intervals for future outcome values vs. inference on policy effects), there are several important differences. First, we rely on permuting residuals, whereas CWZ18 permute blocks of data. Second, we theoretically analyze different types of permutations. In particular, we study the set of all permutations, which yields precise p -values in small samples. This set of permutations cannot be used in the framework of CWZ18 unless the data are iid. Third, we allow for non-stationary data, whereas the prediction methods in CWZ18 are strictly limited to stationary data. Forth, CWZ18 rely on abstract high-level conditions on the test statistics and do not provide any primitive conditions. By contrast, we develop transparent sufficient conditions that can be verified for many traditional and modern CSC methods, and we provide explicit primitive conditions for a large selection of popular approaches. Finally, we establish the validity of our methods with time series data under misspecification and stability, whereas the

theoretical results for weakly dependent data in CWZ18 require correct specification and consistency.

Finally, we show that the problem of making inferences on policy effects can be recast as a structural breaks testing problem with a known break date. Therefore, we build on and contribute to the literature on testing for structural breaks and, in particular, to the literature on structural breaks testing using permutation approaches (e.g., [Antoch and Huskova, 2001](#); [Zeileis and Hothorn, 2013](#)). Besides a different focus (inference on policy effects vs. testing for structural breaks), our paper differs from the existing literature in that we specifically focus on testing at the end of the sample, allow for a very general class of estimators, including modern high-dimensional methods, and provide non-asymptotic performance guarantees under correct specification and misspecification. Our paper is also related to tests for structural breaks at the end of the sample (e.g., [Dufour et al., 1994](#); [Andrews, 2003](#)).² Let us discuss the differences to [Andrews \(2003\)](#)'s end-of-sample instability test based on subsampling in more detail. First, we focus on causal inference, whereas [Andrews \(2003\)](#) is concerned with structural breaks testing. Second, our procedures are exactly valid under exchangeability, and we obtain finite sample bounds under weak conditions on the estimators, while the theoretical properties of [Andrews \(2003\)](#)'s test rely on asymptotic analyses. Third, our methods are valid under misspecification, whereas [Andrews \(2003\)](#) assumes correct specification. Forth, our results under correct specification only require stationarity and weak dependence of $\{u_t\}$, while [Andrews \(2003\)](#)'s test assumes stationarity of the data.³ Finally, our procedures work in conjunction with many modern high-dimensional estimators, whereas [Andrews \(2003\)](#) focuses on low-dimensional GMM models.

Notation.

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. We 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 use the notation $a \asymp b$ to denote $a \lesssim b$ and $b \lesssim a$. For a set A , $|A|$ denotes the cardinality of A . For any $a \in \mathbb{R}$, we define $\lfloor a \rfloor = \max\{z \in \mathbb{Z} : z \leq a\}$ and $\lceil a \rceil = \min\{z \in \mathbb{Z} : z \geq a\}$, where \mathbb{Z} is the set of integers. We use \mathbb{N} to denote the set of natural numbers.

2 A Conformal Inference Method

2.1 The Counterfactual Model

We consider a time series of T outcomes for a treated unit, labeled $j = 1$. During the first T_0 periods, the unit is not treated by a policy and, during the remaining $T - T_0 = T_*$ periods, it is treated by the policy. Extensions to more than one treated unit are discussed in the Appendix. Our typical setting is where T_* is short compared to T_0 . There may be other units that are not exposed to the policy, and they will be introduced below. We denote the observed outcome of the treated unit by Y_{1t} . We employ the potential (latent) outcomes framework (Neyman, 1923; Rubin, 1974) and denote potential outcomes with and without the policy as Y_{1t}^I and Y_{1t}^N . The effect of the policy intervention in period t is $\theta_t = Y_{1t}^I - Y_{1t}^N$.

Our conformal inference method will rely on the following counterfactual modeling framework, which nests many traditional and new methods for counterfactual policy analysis; see Sections 2.3–2.4 for examples.

Assumption 1 (Counterfactual Model). Let $\{P_t^N\}$ be a given sequence of mean-unbiased predictors or proxies for the counterfactual outcomes $\{Y_{1t}^N\}$ in the absence of the policy intervention, that is $\{E(P_t^N)\} = \{E(Y_{1t}^N)\}$. Let $\{\theta_t\}$ be a fixed policy effect sequence with $\theta_t = 0$ for $t \leq T_0$, so that potential outcomes under the intervention are given by $\{Y_{1t}^I\} = \{Y_{1t}^N + \theta_t\}$.⁴ In other words, potential outcomes can be written as

$$\begin{aligned} Y_{it}^N &= P_t^N + u_t \\ Y_{it}^I &= P_t^N + \theta_t + u_t \end{aligned} \quad \Bigg| \quad E(u_t) = 0, \quad t = 1, \dots, T, \quad (CMF)$$

where $\{u_t\}$ is a centered stationary stochastic process. Observed outcomes are related to potential outcomes as $Y_{it} = Y_{it}^N + D_t(Y_{it}^I - Y_{it}^N)$, where $D_t = 1(t > T_0)$.

Assumption 1 introduces the potential outcomes, but also postulates an identifying assumption in the form of the existence of mean-unbiased proxies P_t^N such that $E(P_t^N) = E(Y_{it}^N)$. Assumption 1 allows $\{P_t^N\}$ to be fixed or random and does not impose any restrictions on the dependence between $\{P_t^N\}$ and $\{u_t\}$. In Sections 2.3–2.4, we will discuss specific panel data and time series models that postulate (and identify) what P_t^N is under a variety of conditions. Additional assumptions on the stochastic shock process $\{u_t\}$ will be introduced later, in essence requiring $\{u_t\}$ to be either iid or, more generally, a stationary and weakly dependent process.

Assumption 1 also postulates that the stochastic shock sequence is invariant under the intervention. This is the fundamental identifying assumption. It requires that the timing of the policy intervention is independent of factors that change the distribution of $\{u_t\}$.⁵ If the policy changes the distribution of $\{u_t\}$, one can either interpret our method as a structural breaks test or view the policy effect as random in which case our method yields valid prediction sets; see the Appendix for details.

Often, there is additional information in the form of untreated units, which can serve as controls. Specifically, suppose that there are $J \geq 1$ control units, indexed by $j = 2, \dots, J+1$. We assume that we observe all units for all T periods, although this assumption can be relaxed. Let Y_{jt} denote the observed outcome for these untreated units. This observed outcome is equal to the outcome in the absence of the policy intervention, i.e., $Y_{jt} = Y_{jt}^N$ for $2 \leq j \leq J+1$ and $1 \leq t \leq T$. For each unit, we may also observe a vector of covariates X_{jt} . This motivates a variety of strategies for modeling and identifying P_t^N as discussed below.

2.2 Hypotheses of Interest, Test Statistics, and p -Values

We are interested in testing hypotheses about the trajectory of policy effects in the post-treatment period, $\theta = (\theta_{T_0+1}, \dots, \theta_T)'$. Our main hypothesis of interest is

$$H_0: \theta = \theta^0, \quad (1)$$

where $\theta^0 = (\theta_{T_0+1}^0, \dots, \theta_T^0)'$ is a postulated policy effect trajectory. Hypothesis (1) is a sharp null hypothesis. It fully determines the value of the counterfactual outcome in the absence of the intervention in the post-treatment period since $Y_{1t}^N = Y_{1t}^I - \theta_t = Y_{1t} - \theta_t$. In the Appendix, we show that our method can also be used to test hypotheses about average effects.

To describe our procedure, we write the data under the null hypothesis as $\mathbf{Z} := \mathbf{Z}(\theta^0) = (Z_1, \dots, Z_T)'$, where

$$Z_t = \begin{cases} (Y_{1t}^N, Y_{2t}^N, \dots, Y_{J+1t}^N, X'_{1t}, \dots, X'_{J+1t})', & t \leq T_0 \\ (Y_{1t}^I - \theta_t^0, Y_{2t}^N, \dots, Y_{J+1t}^N, X'_{1t}, \dots, X'_{J+1t})', & t > T_0. \end{cases}$$

Using one of the methods described below, we will obtain a counterfactual proxy estimate, \hat{P}_t^N , based on \mathbf{Z} , and compute the residuals $\hat{u} = (\hat{u}_1, \dots, \hat{u}_T)'$, where $\hat{u}_t = Y_{1t}^N - \hat{P}_t^N$ for $1 \leq t \leq T$. Since P_t^N is computed using $\mathbf{Z} = \mathbf{Z}(\theta^0)$, P_t^N is estimated under the null hypothesis, which is essential for a good small sample performance. In “ideal” settings where the data are iid, imposing the null guarantees the model-free exact finite sample validity of our method; see the Appendix for details. By contrast, when P_t^N is estimated based on the pre-treatment data $\{Z_t\}_{t=1}^{T_0}$ without imposing the null, permuting blocks of residuals does not yield procedures with exact finite sample validity, not even with iid data.

Definition of Test Statistic S. We consider the following test statistic:

$$S(\hat{u}) = S_q(\hat{u}) = \left(\frac{1}{\sqrt{T_*}} \sum_{t=T_0+1}^T |\hat{u}_t|^q \right)^{1/q}.$$

Note that S is constructed such that high values indicate rejection. Different choices of q lead to power against different alternatives. For instance, if the intervention has a large but only temporary effect (i.e., if $|\theta_t|$ is large for few periods), choosing $q = \infty$ yields high power. On the other hand, if the intervention has a permanent effect (i.e., if θ_t is non-zero for many post-treatment periods), tests using S_1 or S_2 exhibit good power properties. In our application, we will be using S_1 , which behaves well under heavy-tailed data. Throughout the paper, when the nature of the statistic is not essential, we write $S = S_q$.

Remark 1 (Choice of Test Statistic). While we focus on S_q , other test statistics can be used as well. For example, when capturing deviations in the average

effect $T_*^{-1} \sum_{t=T_0+1}^T \theta_t$, it is useful to consider $S(\hat{u}) = T_*^{-1/2} \left| \sum_{t=T_0+1}^T \hat{u}_t \right|$. \square

We use (block) permutations to compute p -values. A permutation π is a one-to-one mapping $\pi: \{1, \dots, T\} \mapsto \{1, \dots, T\}$. We denote the set of permutations under study as Π and assume that Π contains the identity map \mathbb{I} . We focus on two different sets of permutations: (i) the set of all permutations, which we call *iid permutations*, Π_{all} , and (ii) the set of all (overlapping) *moving block permutations*, Π_{\rightarrow} .⁶ The elements of Π_{\rightarrow} are indexed by $j \in \{0, 1, \dots, T-1\}$, and the permutation π_j is defined as

$$\pi_j(i) = \begin{cases} i + j & \text{if } i + j \leq T \\ i + j - T & \text{otherwise.} \end{cases}$$

Figure 2 presents a graphical illustration of Π_{all} and Π_{\rightarrow} .

[Figure 2 around here.]

The choice of Π does not matter for the exact finite sample validity of our procedures if the residuals are exchangeable. However, Π_{all} has more elements than Π_{\rightarrow} , allowing us to compute more precise p -values and to test at lower significance levels. For the asymptotic validity under estimator consistency, the

choice of Π depends on the assumptions that we are willing to impose on the stochastic shock sequence $\{u_t\}$ (cf. Section 3.1).

For each $\pi \in \Pi$, let $\hat{u}_\pi = (\hat{u}_{\pi(1)}, \dots, \hat{u}_{\pi(T)})'$ denote the vector of permuted residuals.⁷ The permutation p -value is defined as follows.

Definition of p -Value. The p -value is

$$\hat{p} = 1 - \hat{F}(S(\hat{u})), \text{ where } \hat{F}(x) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \mathbf{1}\{S(\hat{u}_\pi) < x\}. \quad (2)$$

We are often interested in testing pointwise hypotheses about θ_t , $H_0: \theta_t = \theta_t^0$, and in constructing pointwise confidence intervals for θ_t . Pointwise hypotheses can be tested by defining the data under the null as $\mathbf{Z} = (Z_1, \dots, Z_{T_0}, Z_t)'$, provided that P_t^N can be estimated based on \mathbf{Z} . Pointwise $(1 - \alpha)$ confidence intervals for θ_t can be constructed via test inversion as described in Algorithm 1.

Algorithm 1 (Pointwise Confidence Intervals). (i) Choose a fine grid of G candidate values $\tilde{\Theta}_t = \{\tilde{\theta}_t^0, \dots, \tilde{\theta}_t^G\}$. (ii) For $\tilde{\theta}_t^0 \in \tilde{\Theta}_t$, define \mathbf{Z} for the null hypothesis $H_0: \theta_t = \tilde{\theta}_t^0$ and compute the corresponding p -value, $\hat{p}(\tilde{\theta}_t^0)$, using (2). (iii) Return the $(1 - \alpha)$ confidence set $\mathcal{C}_{1-\alpha}(t) = \{\tilde{\theta}_t^0 \in \tilde{\Theta}_t : \hat{p}(\tilde{\theta}_t^0) > \alpha\}$.

2.3 Models for Counterfactual Proxies via Synthetic Control and Panel Data

The availability of control units motivates several strategies for modeling the counterfactual mean proxies P_t^N . We estimate P_t^N based on the imputed data under the null hypothesis, $\mathbf{Z}(\theta^0)$, and write Y_{it}^N instead of $Y_{it}^I - \theta_t^0$ to alleviate the exposition.

2.3.1 Difference-in-Differences Methods

The difference-in-differences method postulates the following model for the counterfactual mean proxy (e.g., [Doudchenko and Imbens, 2016](#), Section 5.1):

$P_t^N = \mu + \frac{1}{J} \sum_{j=2}^{J+1} Y_{jt}^N$. This model automatically embeds the identifying information.

The counterfactual mean proxy can be estimated as

$$\hat{P}_t^N = \frac{1}{T} \sum_{s=1}^T \left(Y_{1s}^N - \frac{1}{J} \sum_{j=2}^{J+1} Y_{js}^N \right) + \frac{1}{J} \sum_{j=2}^{J+1} Y_{jt}^N.$$

2.3.2 Synthetic Control and Constrained Lasso

The canonical SC method (e.g., [Abadie and Gardeazabal, 2003](#); [Abadie et al., 2010, 2015](#)) postulates the following model:

$$P_t^N = \sum_{j=2}^{J+1} w_j Y_{jt}^N, \text{ where } w \geq 0 \text{ and } \sum_{j=2}^{J+1} w_j = 1. \quad (3)$$

We need to impose an identification condition that allows us to identify the weights w , for example:⁸

- Assume that the structural shocks u_t for the treated unit are uncorrelated with contemporaneous values of the outcomes, namely:
 $E(u_t Y_{jt}^N) = 0$ for $2 \leq j \leq J+1$.

The counterfactual is estimated as $\hat{P}_t^N = \sum_{j=2}^{J+1} \hat{w}_j Y_{jt}^N$. We focus on the following canonical SC estimator for w .⁹

$$\hat{w} = \arg \min_w \sum_{t=1}^T \left(Y_{1t}^N - \sum_{j=2}^{J+1} w_j Y_{jt}^N \right)^2 \text{ s.t. } w \geq 0 \text{ and } \sum_{j=2}^{J+1} w_j = 1. \quad (4)$$

As an alternative, we can consider the more flexible model¹⁰

$$P_t^N = \mu + \sum_{j=2}^{J+1} w_j Y_{jt}^N, \text{ where } \|w\|_1 \leq 1, \quad (5)$$

maintaining the same identifying assumption (SC). The counterfactual is estimated as $\hat{P}_t^N = \hat{\mu} + \sum_{j=2}^{J+1} \hat{w}_j Y_{jt}^N$ by the ℓ_1 -constrained least squares estimator, or constrained Lasso (e.g., [Raskutti et al., 2011](#)):

$$(\hat{\mu}, \hat{w}) = \arg \min_{(\mu, w)} \sum_{t=1}^T \left(Y_{1t}^N - \mu - \sum_{j=2}^{J+1} w_j Y_{jt}^N \right)^2 \quad \text{s.t. } \|w\|_1 \leq 1. \quad (6)$$

The advantage over other penalized regression methods discussed next is that constrained Lasso is essentially tuning-free, does not rely on any sparsity conditions, and is valid for dependent data under weak assumptions. Moreover, constrained Lasso encompasses both difference-in-differences and canonical SC as special cases (by setting $w = (1/J, \dots, 1/J)'$ and $\mu = 0, w \geq 0$, respectively) and, thus, provides a unifying approach for the regression-based estimation of P_t^N .

Section 4.2 provides primitive conditions that guarantee that the SC and the constrained Lasso estimators are valid in our framework in settings with potentially many control units (large J). Finally, we note that it is straightforward to incorporate (transformations of) covariates X_{jt} into the estimation problems (4) and (6).

2.3.3 Penalized Regression Methods

Consider a linear model for P_t^N : $P_t^N = \mu + \sum_{j=2}^{J+1} w_j Y_{jt}^N$. We maintain the identifying

assumption (SC). The counterfactual is estimated by $\hat{P}_t^N = \hat{\mu} + \sum_{j=2}^{J+1} \hat{w}_j Y_{jt}^N$, where

$$(\hat{\mu}, \hat{w}) = \arg \min_{(\mu, w)} \sum_{t=1}^T \left(Y_{1t}^N - \mu - \sum_{j=2}^{J+1} w_j Y_{jt}^N \right)^2 + \mathcal{P}(w), \quad (7)$$

and $\mathcal{P}(w)$ is a penalty function that penalizes deviations away from zero. If it is desired to penalize deviations away from other focal points w^0 , for example, $w^0 = (1/J, \dots, 1/J)$ used in the difference-in-differences approach, we may always

use instead: $\mathcal{P}(w) \leftarrow \mathcal{P}(w - w^0)$. Note that it is straightforward to incorporate covariates X_{jt} into the estimation problem (7).

Different variants of $\mathcal{P}(w)$ can be considered. Examples include: Lasso (Tibshirani, 1996), where $\mathcal{P}(w) = \lambda \|w\|_1$ and λ is a tuning parameter; Elastic Net (Zou and Hastie, 2005), where $\mathcal{P}(w) = \lambda \left((1 - \alpha) \|w\|_2^2 + \alpha \|w\|_1 \right)$ and λ and α are tuning parameters; Lava (Chernozhukov et al., 2017), where $\mathcal{P}(w) = \inf_{a+b=w} \lambda \left((1 - \alpha) \|a\|_2^2 + \alpha \|b\|_1 \right)$ and λ and α are tuning parameters.

In the context of CSC methods, Lasso was used by Valero (2015), Li and Bell (2017), and Carvalho et al. (2018), while Doudchenko and Imbens (2016) proposed to use Elastic Net. We will impose only weak requirements on the performance of the estimators (pointwise consistency and consistency in prediction norm), which implies that these estimators are valid in our framework under any set of sufficient conditions that exists in the literature.

2.3.4 Interactive Fixed Effects, Factor, and Matrix Completion Models

Consider the following interactive FE model for treated and untreated units:

$$Y_{jt}^N = \lambda_j' F_t + X_{jt}' \beta + u_{jt}, \quad \text{for } 1 \leq j \leq J+1 \text{ and } 1 \leq t \leq T,$$

where F_t are unobserved factors, λ_j are unit-specific factor loadings, and β is a vector of common coefficients. Model (8) nests the classical factor model when $\beta = 0$ and also covers the traditional linear FE model, in which $\lambda_j' F_t = \lambda_j + F_t$. Consider the following assumption.

(FE) Assume that u_{jt} is uncorrelated with (X_{jt}, F_t, λ_j) , as well as other identification conditions in Bai (2009).

The model leads to the following proxy:

$$P_t^N = \lambda_1' F_t + X_{1t}' \beta. \quad (9)$$

Counterfactual proxies are estimated by $\hat{P}_t^N = \hat{\lambda}_1' \hat{F}_t + X_{1t}' \hat{\beta}$, where $\hat{\lambda}_1$ and \hat{F}_t , and $\hat{\beta}$ are obtained using the alternating least squares method applied to the model (8); see, for example, [Bai \(2009\)](#) and [Hansen and Liao \(2019\)](#) for a version with high-dimensional covariates.

[Hsiao et al. \(2012\)](#) appears to be the first work that proposed the use of factor models for predicting the (missing) counterfactual responses specifically in SC settings. [Gobillon and Magnac \(2016\)](#) and [Xu \(2017\)](#) employ [Bai \(2009\)](#)'s estimator in this setting, albeit provide no formally justified inference methods. Formal inference results for interactive FE and factor models in SC designs are developed in [Chan and Kwok \(2016\)](#) and [Li \(2018\)](#) among others.¹¹

Other recent applications to predicting counterfactual responses include [Amjad et al. \(2018\)](#) and [Athey et al. \(2018\)](#) (using, respectively, singular value thresholding and the nuclear norm penalization).¹² Our method delivers a way to perform valid inference for policy effects using any of the factor model estimators used in these proposals applied to the complete data under the null.¹³ We shall be focusing on [Bai \(2009\)](#)'s alternating least squares estimator¹⁴ and on matrix completion via nuclear norm penalization when verifying our conditions.

2.4 Models for Counterfactual Proxies via Time Series and Fused Models

2.4.1 Simple Time Series Models

If no control units are available, one can use time series models for the single unit exposed to the intervention. For example, consider the following autoregressive model:¹⁵

$$\begin{aligned} Y_{1t}^N - \mu &= \rho(Y_{1(t-1)}^N - \mu) + u_t \\ Y_{1t}^I - \mu &= \rho(Y_{1(t-1)}^N - \mu) + \theta_t + u_t \end{aligned} \quad \left| \quad E(u_t) = 0, \quad \{u_t\} \text{ iid}, \quad t = 1, \dots, T. \quad (10)$$

In model (10), the mean unbiased proxy is given by $P_t^N = \mu + \rho(Y_{1(t-1)}^N - \mu)$. Note that the policy effect here is transitory, namely it does not feed-forward itself on

the future values of Y_{it}^I beyond the current values.¹⁶ Under the null hypothesis, we can impute the unobserved counterfactual as $Y_{it}^N = Y_{it} - \theta_t$ and estimate the model using traditional time series methods, and we can conduct inference by permuting the residuals.

The simplest form of the autoregressive model is the AR(K) process, where the $\rho(\cdot)$ take the form: $\rho(\cdot) = \sum_{k=0}^K \rho_k L^k(\cdot)$, where L is the lag operator. There are many identifying conditions for these models, see, for example, [Hamilton \(1994\)](#) or [Brockwell and Davis \(2013\)](#). More generally, we can use a nonlinear function of lag operators, $\rho(\cdot) = m(\cdot, L^1(\cdot), \dots, L^k(\cdot))$, as, for example, when applying neural networks to time series data (e.g., [Chen and White, 1999](#); [Chen et al., 2001](#)), and we refer to the latter for identifying conditions.

2.4.2 Fused Time-Series/Panel Models

A simple and generic way to combine the insights from the panel data and time series models is as follows. Consider the system of equations:

$$\begin{aligned} Y_{it}^N &= C_{it}^N + \varepsilon_t \\ Y_{it}^I &= C_{it}^N + \theta_t + \varepsilon_t \end{aligned} \quad \left| \begin{array}{l} \varepsilon_t = \rho(\varepsilon_{t-1}) + u_t, \{u_t\} \text{ iid}, E(u_t) = 0, \\ \{u_t\} \text{ is independent of } \{C_{it}^N\}, \end{array} \right| \quad t = 1, \dots, T, \quad (11)$$

where C_{it}^N is a panel model proxy for Y_{it}^N , identified by one of the panel data methods. Note that the model has the autoregressive formulation:

$$Y_{it}^N = C_{it}^N + \rho(Y_{i(t-1)}^N - C_{i(t-1)}^N) + u_t, \text{ thereby generalizing the previous model.}$$

Here the mean unbiased proxy for Y_{it}^N is given by $P_{it}^N = C_{it}^N + \rho(\varepsilon_{t-1})$. P_{it}^N is a better proxy than C_{it}^N because it provides an additional noise reduction through prediction of the stochastic shock by its lag. The model combines any favorite panel model C_{it}^N for counterfactuals with a time series model for the stochastic shock model in a nice way: we can identify C_{it}^N under the null by ignoring the time series structure, and then identify the time series structure of the residuals

$Y_{1t}^N - C_t^N$. Estimation can proceed analogously. This approach will often improve the size accuracy of our inferential procedures.

3 Theory

When the data are iid (or exchangeable), our procedure is exactly valid in finite samples as shown in the Appendix. In this section, we establish the validity of our inference methods with time series data. Our results are non-asymptotic in nature and, hence, hold in *finite samples*. Finite sample bounds are provided for the size properties of our procedure; these bounds imply that our approach is exact as $T_0 \rightarrow \infty$. In Section 3.1, we establish the validity of our procedure when the estimator of P_t^N satisfies weak and easy-to-verify small error conditions (pointwise consistency and consistency in the prediction norm). This result accommodates non-stationary data and only requires stationarity and weak dependence of the stochastic shock process $\{u_t\}$. In Section 3.2, we consider a setting that accommodates misspecification and inconsistent estimators. We show that if the data are stationary and weakly dependent, our procedure is valid, provided that the estimators are stable.

3.1 Approximate Validity under Estimator Consistency

The main condition underlying the results in this section is the following assumption on the stochastic shock process.

Assumption 2 (Regularity of the Stochastic Shock Process). *Assume that the density function of $S(u)$ exists and is bounded, and that the stochastic process $\{u_t\}_{t=1}^T$ satisfies one of the following conditions.*

1. $\{u_t\}_{t=1}^T$ are iid, or
2. $\{u_t\}_{t=1}^T$ are stationary, strongly mixing, with sum of mixing coefficient bounded by M .

Assumption 2 allows the data to be non-stationary and exhibit general dependence patterns. Assumption 2.1 of iid shocks is our first sufficient condition. Under this condition, we will be able to use iid permutations, giving us a precise estimate of the p -value. The iid assumption can be replaced by Assumption 2.2, which holds for many commonly encountered stochastic processes such as ARMA and GARCH. It can be easily replaced by an even weaker ergodicity condition, as can be inspected in the proofs. Under this assumption, we will have to rely on the moving block permutations.

Remark 2 (Heteroscedasticity). Assumption 2 does not rule out conditional heteroscedasticity in the stochastic shock process $\{u_t\}$. Unconditional heteroscedasticity is allowed in $\{Z_t\}$ but not in $\{u_t\}$. When we suspect unconditional heteroscedasticity in $\{u_t\}$, we can apply another filter or model to obtain “standardized residuals” from $\{\hat{u}_t\}$. This will generally require another layer of modeling assumptions, leading to an overall procedure that reduces the data to “fundamental” shocks that are assumed to be stationary under the null. \square

We also impose the following condition on the estimation error under the null hypothesis. Let $P^N = (P_1^N, \dots, P_T^N)'$ and $\hat{P}^N = (\hat{P}_1^N, \dots, \hat{P}_T^N)'$.

Assumption 3 (Consistency of the Counterfactual Estimators under the Null). *Let there be sequences of constants δ_T and γ_T converging to zero. Assume that with probability $1 - \gamma_T$,*

1. *the mean squared estimation error is small, $\|\hat{P}^N - P^N\|_2^2 / T \leq \delta_T^2$;*
2. *for $T_0 + 1 \leq t \leq T$, the pointwise errors are small, $|\hat{P}_t^N - P_t^N| \leq \delta_T$.*

Assumption 3 imposes weak and easy-to-verify conditions on the performance of the estimators \hat{P}_t^N of the counterfactual mean proxies P_t^N . These conditions are readily implied by the existing results for many estimators discussed in Section 2.

In Section 4, we provide explicit primitive conditions and references to primitive conditions implying Assumption 3.¹⁷

Theorem 1 (*Approximate Validity under Consistent Estimation*). *Assume that T_* is fixed. Suppose that Assumptions 1 and 3 hold. Impose Assumption 2.1 if $\Pi = \Pi_{\text{all}}$; impose Assumption 2.2 if $\Pi = \Pi_{\rightarrow}$. Assume $S(u)$ has a density function bounded by D under the null. Then, under the null hypothesis, the p -value is approximately unbiased in size:*

$$|P(\hat{p} \leq \alpha) - \alpha| \leq C(\tilde{\delta}_T + \delta_T + \sqrt{\delta_T} + \gamma_T),$$

where $\tilde{\delta}_T = (T_*/T_0)^{1/4}(\log T)$ and the constant C depends on T_* , M and D , but not on T .

The above bound is non-asymptotic, allowing us to claim uniform validity with respect to a rich variety of data generating processes. Using simulations and empirical examples, we verify that our tests have good power and generate meaningful empirical results. There are other considerations that also affect power. For example, the better the model for P_t^N , the less variance the stochastic shocks will have, subject to assumed invariance to the policy. The smaller the variance of the shocks, the more powerful the testing procedure will be.

3.2 Approximate Validity under Estimator Stability

Misspecification is an important practical concern, and consistency of the estimators of the counterfactual mean proxies P_t^N may be questionable in certain settings. The classical analysis of misspecification focuses on convergence to pseudo-true values (e.g., [White, 1996](#)). If it is possible to show that the estimator of the counterfactual mean proxy, \hat{P}_t^N , is consistent for some pseudo-true value P_t^{N*} and that $\{Y_{it}^N - P_t^{N*}\}_{t=1}^T$ is stationary and weakly dependent, the theoretical results in Section 3.1 imply the validity of our procedure. Pseudo-true consistency can often be verified for low-dimensional models, but consistency results under misspecification remain elusive in high-dimensional settings.

Therefore, we consider a notion of approximate exchangeability, which only requires the estimator to be *stable* instead of consistent for a pseudo-true value. This stability condition does not require \hat{P}_t^N to be consistent for anything, nor does it rely on correct specification of the counterfactual mean proxies. In the Appendix, we illustrate the difference between consistency and stability based on the analytically tractable example of Ridge regression.

The basic idea underlying the theoretical analysis here is as follows. If the estimators are non-random or independent of the data, then stationarity and weak dependence of the data would mean that \hat{p} based on moving block permutations approximately has a uniform distribution under the null. This result follows from uniform laws of large numbers for dependent data. However, in practice, the estimators are computed using the data and are thus not independent of the data. Our key insight is that stable estimators are approximately independent of individual observations.

We now formalize the notion of stability of an estimator. To emphasize the dependence of $S(\hat{u})$ on the estimator, with a slight abuse of notation, we write $S(\mathbf{Z}, \beta) = \phi(Z_{T_0+1}, \dots, Z_{T_0+T_*}; \beta)$. Let $\{\tilde{Z}_t\}_{t=1}^T$ be iid from the distribution of Z_1 and independent of \mathbf{Z} . For any $H \subset \{1, \dots, T\}$, let $Z_{t,H} = Z_t \mathbf{1}\{t \notin H\} + \tilde{Z}_t \mathbf{1}\{t \in H\}$, and $\mathbf{Z}_H = \{Z_{t,H}\}_{t=1}^T$. Hence, \mathbf{Z}_H is a perturbed version of \mathbf{Z} under H , i.e., \mathbf{Z} with elements in H replaced by $\{\tilde{Z}_t\}_{t \in H}$.

By stability, we mean that the estimator computed using \mathbf{Z} is similar to that computed using \mathbf{Z}_H for $H \in \mathbb{H}$. Let $R \in \mathbb{N}$ and define $m = \lfloor T_0 / R \rfloor$. The class $\mathbb{H} = \{\tilde{H}_1, \dots, \tilde{H}_R\}$ contains R members with $|\tilde{H}_j| \leq 3m$ elements. The plan is to require stability under $R \asymp T_0 / \log(T_0)$ (so $|\tilde{H}_j| \asymp \log(T_0)$). Since $\log(T_0) \ll T_0$, swapping out $O(\log(T_0))$ out of $T_0 + T_*$ data points should not cause a large change in the estimator for reasonable estimators.

We now give precise definitions of sets in \mathbb{H} . For $j \in \{1, \dots, R\}$, let $H_j = \{(j-1)m+1, \dots, jm\}$. Since the test statistic depends on T_* data points after obtaining the estimator, defining \mathbb{H} to be $\{H_1, \dots, H_R\}$ is not enough for technical arguments; we need a “wedge” to ensure that these T_* data points do not cause a problem. To do so, we enlarge H_j as follows. Let $k \in \mathbb{N}$ satisfy $T_* < k < m$. We let \tilde{H}_j denote the k -enlargement of H_j , i.e., $\tilde{H}_j = \{s : \min_{t \in H_j} |s - t| \leq k\}$. Note that $\tilde{H}_j = \{(j-1)m+1-k, \dots, jm+k\}$ for $2 \leq j \leq R-1$, $\tilde{H}_1 = \{1, \dots, m+k\}$ and $\tilde{H}_R = \{(R-1)m+1-k, \min\{Rm+k, T\}\}$.

Assumption 4 (Estimator Stability). *Let $\Pi = \Pi_{\rightarrow}$. There exist non-decreasing functions $\varrho_T(\cdot)$ such that $P\left(\max_{\pi \in \Pi} \left|S\left(\mathbf{Z}^\pi, \hat{\beta}(\mathbf{Z})\right) - S\left(\mathbf{Z}^\pi, \hat{\beta}(\mathbf{Z}_H)\right)\right| \leq \varrho_T(|H|)\right) \geq 1 - \gamma_{1,T}$ and*

$$P\left(\max_{\pi \in \Pi} \left|S\left(\dot{\mathbf{Z}}^\pi, \hat{\beta}(\mathbf{Z})\right) - S\left(\dot{\mathbf{Z}}^\pi, \hat{\beta}(\mathbf{Z}_H)\right)\right| \leq \varrho_T(|H|)\right) \geq 1 - \gamma_{1,T} \text{ for any}$$

$H \in \{\tilde{H}_1, \dots, \tilde{H}_R\}$, where $\dot{\mathbf{Z}} = \mathbf{Z}$ and $\dot{\mathbf{Z}}$ is independent of $(\mathbf{Z}, \{\tilde{\mathbf{Z}}_i\}_{i=1}^T)$.

Assumption 4 specifies the estimator stability condition. It strengthens the perturb-one sensitivity of (Lei et al., 2018, Assumption A.3). When the model is misspecified, Assumption 4 holds whenever the estimator $\hat{\beta}(\mathbf{Z})$ is consistent to a pseudo-true parameter value. However, it is more general in that the estimator $\hat{\beta}(\mathbf{Z})$ need not converge to any non-random quantity as long as it is stable under perturbations in a few observations. This feature is crucial in our setting as it allows us to accommodate high-dimensional CSC methods for many of which consistency results under misspecification are not available. Primitive sufficient conditions for Assumption 4 are provided in the Appendix.

Let $\Psi(x; \beta) = P(\phi(Z_{T_0+1}, \dots, Z_{T_0+T_*}; \beta) \leq x)$. Our strategy is to show that, under the null hypothesis, $\hat{F}(\phi(Z_{T_0+1}, \dots, Z_{T_0+T_*}; \hat{\beta}(\mathbf{Z})))$ is approximately uniform on $(0, 1)$. We exploit the stability condition in Assumption 4 and show that

$\hat{F}(\phi(Z_{T_0+1}, \dots, Z_{T_0+T_*}; \hat{\beta}(\mathbf{Z})))$ can be approximated by

$\Psi\left(\phi(\bar{Z}_{T_0+1}, \dots, \bar{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\tilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\tilde{H}_R})\right)$, which has the uniform distribution on $(0, 1)$. Here $(\bar{Z}_{T_0+1}, \dots, \bar{Z}_{T_0+T_*})$ has the same distribution as $(Z_{T_0+1}, \dots, Z_{T_0+T_*})$ and is independent of $\mathbf{Z}_{\tilde{H}_R}$. This essentially confirms the above intuition that for stable estimators, $\hat{\beta}(\mathbf{Z})$ is almost independent of the last few observations $(Z_{T_0+1}, \dots, Z_{T_0+T_*})$.

We impose the following regularity conditions on the data.

Assumption 5 (Regularity of the Data). *The data under the null, $\{Z_t\}_{t=1}^T$, are stationary and β -mixing with coefficient $\beta_{\text{mixing}}(\cdot)$ satisfying $\beta_{\text{mixing}}(i) \leq D_1 \exp(-D_2 i^{D_3})$ for some constants $D_1, D_2, D_3 > 0$. For $1 \leq j \leq R$, there exist sequences $\xi_T > 0$ and $\gamma_{2,T} = o(1)$ such that*

$$P\left(\sup_{x \in \mathbb{R}} \left| \partial \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\tilde{H}_j})\right) / \partial x \right| \leq \xi_T\right) \geq 1 - \gamma_{2,T}.$$

Stationarity and β -mixing are commonly imposed conditions on time series data. For a large class of Markov chains, GARCH and various stochastic volatility models, $D_3 = 1$ (cf. Carrasco and Chen, 2002). Let $(\dot{Z}_{T_0+1}, \dots, \dot{Z}_{T_0+T_*})$ be an independent copy of $(Z_{T_0+1}, \dots, Z_{T_0+T_*})$ and also independent of $(\mathbf{Z}, \{\tilde{Z}_t\}_{t=1}^T)$. The bounded derivative of $\Psi\left(x; \hat{\beta}(\mathbf{Z}_{\tilde{H}_j})\right)$ condition says that the density of $\phi(\dot{Z}_t, \dots, \dot{Z}_{t+T_*-1}; \hat{\beta}(\mathbf{Z}_{\tilde{H}_j}))$ conditional on $\hat{\beta}(\mathbf{Z}_{\tilde{H}_j})$ is bounded by ξ_T with high probability. The bounded density condition states that the distribution of the residual does not collapse into a degenerate one or one with point mass. In many cases, $\xi_T = O(1)$ for continuous distributions. For example, if (Y_t, X_t) is jointly Gaussian and the variance of Y_t given X_t is bounded below by a constant, then for any w satisfying the SC restrictions, the density of $Y_t - X_t'w$ is bounded by a constant that does not depend on w .

The following result states the approximate validity of our testing procedure.

Theorem 2 (Approximate Validity under Estimator Stability). *Let $\Pi = \Pi_{\rightarrow}$.*

Suppose that Assumptions 4 and 5 hold. Then, under the null hypothesis, there exists a constant $C_1 > 0$ depending only on D_1 , D_2 and D_3 such that for any R with $k < \lfloor T_0 / R \rfloor$ and $R < T_0 / 2$,

$$\begin{aligned} |P(\hat{p} \leq \alpha) - \alpha| &\leq C_1 \sqrt{\xi_T \varrho_T (T_0 / R + 2k)} + C_1 \left(T_0^{-1} R [\log(T_0 / R)]^{1/D_3} \right)^{1/4} \\ &+ C_1 \exp\left(-(k - T_* + 1)^{1/D_3}\right) + C_1 \sqrt{\gamma_{1,T}} + C_1 \sqrt{\gamma_{2,T}}. \end{aligned}$$

In the theoretical arguments, we actually show a stronger result. The above bound holds for $E | P(\hat{p} \leq \alpha | \hat{\beta}(\mathbf{Z}_{\tilde{H}_R})) - \alpha |$. Since the stability condition states that $\hat{\beta}(\mathbf{Z}_{\tilde{H}_R}) \approx \hat{\beta}(\mathbf{Z})$, this means that \hat{p} conditional on $\hat{\beta}(\mathbf{Z})$ almost has a uniform distribution on $(0, 1)$; with iid or exchangeable data, \hat{p} conditional on $\hat{\beta}(\mathbf{Z})$ has an exact uniform distribution. Therefore, we can view Theorem 2 as a result for approximate exchangeability.

Due to the exponential decay of $\beta_{\text{mixing}}(\cdot)$, the bound in Theorem 2 tends to zero if we choose k to be a slowly growing sequence and T_0 / R to be of the same order. For example, we can choose k and R such that $k \asymp T_0 / R \asymp \log T_0$. Since $|\tilde{H}_j| = \lfloor T_0 / R \rfloor + 2k$, Assumption 5 only requires that the changes to $S(\mathbf{Z}^\pi, \hat{\beta}(\mathbf{Z}))$ are small if we replace only $\log T_0$ observations in computing $\hat{\beta}(\mathbf{Z})$. Under finite dependence, it suffices to choose k and T_0 / R to be large enough constants. Note that R is only needed in the theoretical arguments; we do not need to choose R when implementing the proposed procedure.

The theoretical analysis in this section suggests that allowing for both unrestricted patterns of non-stationarity and misspecification is not possible in general. To obtain valid inferences with non-stationary data, one has to either rely on correct specification and consistency or impose assumptions on the particular structure of the non-stationarity, which allow for pre-processing the data to make them stationary.

4 Sufficient Conditions for Consistent Estimation

In this section, we revisit the representative models of counterfactual proxies introduced in Section 2. Primitive conditions are provided to guarantee that the estimation of the counterfactual mean proxies is accurate enough for the asymptotic validity of the proposed procedure. In particular, these conditions can be used to verify Assumption 3. The regularity conditions (e.g., bounded moments, weak serial dependence) for different models are stated in the Appendix and are commonly imposed in the literature for these models. The counterfactual mean proxies P_t^N are estimated based on the imputed data under the null, $\mathbf{Z}(\theta^0)$, and we write Y_{it}^N instead of $Y_{it}^I - \theta_t^0$ to alleviate the exposition. All the results in this section assume that $T_0 \rightarrow \infty$ and $J \rightarrow \infty$ (if J is present in the model).

4.1 Difference-in-Differences

In Section 2.3.1, we have seen that the counterfactual mean proxies implied by the canonical difference-in-differences model are: $P_t^N = \mu + J^{-1} \sum_{j=2}^{J+1} Y_{jt}^N$. We

consider the following estimator: $\hat{P}_t^N = \hat{\mu} + \frac{1}{J} \sum_{j=2}^{J+1} Y_{jt}$, where

$$\hat{\mu} = \frac{1}{T} \sum_{t=1}^T \left(Y_{1t}^N - \frac{1}{J} \sum_{j=2}^{J+1} Y_{jt}^N \right) = \mu + \frac{1}{T} \sum_{t=1}^T u_t. \text{ Since } \hat{P}_t^N - P_t^N = \hat{\mu} - \mu, \text{ Assumption 3 holds}$$

for the simple difference-in-differences model provided that $T^{-1} \sum_{t=1}^T u_t = o_p(1)$,

which is true under very weak conditions.

4.2 Synthetic Control and Constrained Lasso

Several models in Section 2 (including SC and constrained Lasso) imply a structure in which the counterfactual proxy is a linear function of observed outcomes of untreated units.

To provide a unified framework for these models, we use Y to denote a generic vector of outcomes and X to denote the design matrix throughout this section.

For example, in Section 2, we set $Y = Y_1^N$ and $X = (Y_2^N, \dots, Y_{J+1}^N)$, where $Y_j^N = (Y_{j1}^N, \dots, Y_{jT}^N)' \in \mathbb{R}^T$ for $1 \leq j \leq J+1$. These models can be written as

$$Y = Xw + u, \quad (12)$$

where $u = (u_1, \dots, u_T)' \in \mathbb{R}^T$. Identification is achieved by requiring that X and u be uncorrelated (cf. Condition (SC)).

Under the framework in (12), different models correspond to different specifications for the weight vector w . For the SC model in Section 2.3.2, w is an unknown vector whose elements are nonnegative and sum up to one. More generally, one can simply restrict w to be any vector with bounded ℓ_1 -norm. This is the constrained Lasso estimator.

Since P_t^N is the t -th element of the vector Xw , the natural estimator is \hat{P}_t^N being the t -th element of $X\hat{w}$, where \hat{w} is an estimator for w . The estimation of w depends on the specification. Let \mathcal{W} be the parameter space for w . We consider the following version of the original SC estimator

$$\hat{w} = \arg \min_w \|Y - Xw\|_2 \quad \text{s.t. } w \in \mathcal{W} = \{v \geq 0, \|v\|_1 = 1\}. \quad (13)$$

The constrained Lasso estimator is

$$\hat{w} = \arg \min_w \|Y - Xw\|_2 \quad \text{s.t. } w \in \mathcal{W} = \{v : \|v\|_1 \leq K\}, \quad (14)$$

where K is bounded and $K > 0$. In light of the estimator (13), a natural choice is $K = 1$.

In general, we choose the parameter space \mathcal{W} to be an arbitrary subset of an ℓ_1 -ball with bounded radius. The following result gives very mild conditions under which the constrained least squares estimators are consistent and satisfy Assumption 3.¹⁸

Lemma 1 (Constrained Least Squares Estimators). *Consider*

$\hat{w} = \arg \min_v \|Y - Xv\|_2$ s.t. $v \in \mathcal{W}$, where \mathcal{W} is a subset of $\{v : \|v\|_1 \leq K\}$ and K is bounded. Assume $w \in \mathcal{W}$, the data are β -mixing with exponential speed, and other assumptions listed at the beginning of the proof, including the identification condition (SC), then the estimator enjoys the performance bounds stated in the proof, in particular: $\frac{1}{T} \sum_{t=1}^T (\hat{P}_t^N - P_t^N)^2 = o_p(1)$ and $\hat{P}_t^N - P_t^N = o_p(1)$, for any $T_0 + 1 \leq t \leq T$.

Lemma 1 provides several features that are important for counterfactual inference in our setup. First, we allow J to be large relative to T . To be precise, we only require $\log J = o(T^c)$, where $c > 0$ is a constant depending only on the β -mixing coefficients; see the Appendix for details. This is particularly relevant for settings in which the number of (potential) control units and the number of time periods have a similar order of magnitude as in our empirical application in Section 5. Second, Lemma 1 does not rely on any sparsity assumptions on w , allowing for dense vectors. Third, compared to typical high-dimensional estimators (e.g., Lasso or Dantzig selector), our estimator does rely on tuning parameters that can be difficult to choose in times series settings. Finally, Lemma 1 provides new theoretical consistency results for the canonical SC estimator in settings with time series data and potentially very many control units.

4.3 Models with Factor Structures

The models for counterfactual proxies introduced in Section 2.3.4 have factor structures. We provide estimation results for pure factor models (without regressors), factor models with regressors (interactive FE models), and matrix completion models. In this subsection, following standard notation, we let $N = J + 1$.

4.3.1 Pure Factor Models

Recall from Section 2.3.4 the standard factor model $Y_{jt}^N = \lambda_j' F_t + u_{jt}$, where $F = (F_1, \dots, F_T)' \in \mathbb{R}^{T \times k}$ and $\Lambda = (\lambda_1, \dots, \lambda_N)' \in \mathbb{R}^{N \times k}$ represent the k -dimensional unobserved factors and their loadings, respectively. The counterfactual proxy for Y_{jt}^N is $P_t^N = \lambda_1' F_t$. We identify P_t^N by imposing the condition that the idiosyncratic terms and the factor structure are uncorrelated (cf. Condition (FE)).

We use the standard principal component analysis (PCA) for estimating P_t^N .¹⁹ Let $Y^N \in \mathbb{R}^{T \times N}$ be the matrix whose (t, j) entry is Y_{jt}^N . We compute $\hat{F} = (\hat{F}_1, \dots, \hat{F}_T)' \in \mathbb{R}^{T \times k}$ to be the matrix containing the eigenvectors corresponding to the largest k eigenvalues of $Y^N (Y^N)'$ with $\hat{F}' \hat{F} / T = I_k$. Let $\hat{\lambda}_j'$ denote the j -th row of $\hat{\Lambda} = (Y^N)' \hat{F} / T$. Let \hat{F}_t' denote the t -th row of \hat{F} . Our estimate for P_t^N is $\hat{P}_t^N = \hat{\lambda}_1' \hat{F}_t$. The following lemma guarantees the validity of this estimator in our context under mild regularity conditions.

Lemma 2 (Pure Factor Model). *Assume standard regularity conditions given in Bai (2003), including the identification condition (FE). Consider the factor model and the principal component estimator. Then, for any $1 \leq t \leq T$, as $N \rightarrow \infty$ and $T \rightarrow \infty$, we have $\hat{P}_t^N - P_t^N = O_p(1/\sqrt{N} + 1/\sqrt{T})$ and*

$$\frac{1}{T} \sum_{t=1}^T (\hat{P}_t^N - P_t^N)^2 = O_p(1/N + 1/T).$$

The only requirement on the sample size is that both N and T need to be large. Similar to Theorem 3 of Bai (2003), we do not restrict the relationship between N and T . This is flexible enough for a wide range of applications in practice as the number of units is allowed to be much larger than, much smaller than, or similar to the number of time periods.

4.3.2 Factor plus Regression Model: Interactive FE Model

Now we study the general form of panel models with interactive FEs. Following Section 2.3.4, these models take the form $Y_{jt}^N = \lambda_j' F_t + X_{jt}' \beta + u_{jt}$, where $X_{jt} \in \mathbb{R}^{k_x}$ are observed covariates and $F = (F_1, \dots, F_T)' \in \mathbb{R}^{T \times k}$ and

$\Lambda = (\lambda_1, \dots, \lambda_N)' \in \mathbb{R}^{N \times k}$ represent the k -dimensional unobserved factors and their loadings, respectively. The counterfactual proxy for Y_{it}^N is $P_t^N = \lambda_1' F_t + X_{it}' \beta$. In this model, we identify the counterfactual proxy through the condition that the idiosyncratic terms are independent of the factor structure and the observed covariates (cf. Condition (FE)).

The two most popular estimators are the common correlated effects (CCE) estimator by [Pesaran \(2006\)](#) and the iterative least squares estimator by [Bai \(2009\)](#). We focus on the iterative least squares approach, but analogous results can be established for CCE estimators. The notations for F_t , λ_j , \hat{F}_t and $\hat{\lambda}_j$ are the same as before. We compute

$$(\hat{F}, \hat{\Lambda}, \hat{\beta}) = \arg \min_{F, \Lambda, \beta} \sum_{t=1}^T \sum_{j=1}^N (Y_{jt}^N - X_{jt}' \beta - F_t' \lambda_j)^2 \quad \text{s.t.} \quad F'F/T = I_k, \quad \Lambda' \Lambda = \text{Diagonal}_k.$$

The estimate for P_t^N is $\hat{P}_t^N = \hat{\lambda}_1' \hat{F}_t + X_{it}' \hat{\beta}$. The following result states the validity of applying this estimator in conjunction with our inference method.

Lemma 3 (Interactive FE Model). *Assume the standard conditions in [Bai \(2009\)](#), including the identification condition (FE). Then, for any $1 \leq t \leq T$,*

$$\hat{P}_t^N - P_t^N = O_p(1/\sqrt{T} + 1/\sqrt{N}) \quad \text{and} \quad \frac{1}{T} \sum_{t=1}^T (\hat{P}_t^N - P_t^N)^2 = O_p(1/T + 1/N).$$

Under the conditions in Theorem 3 of [Bai \(2009\)](#), N is of the same order as T so that rate is really $T^{-1/2}$; however, the stated bound should hold more generally.

4.3.3 Matrix Completion via Nuclear Norm Regularization

Suppose that

$$Y_{jt}^N = M_{jt} + u_{jt}, \quad \text{for } 1 \leq j \leq J+1 \text{ and } 1 \leq t \leq T, \quad (15)$$

where M_{jt} is the (j, t) -element of an unknown matrix $M \in \mathbb{R}^{(J+1) \times T}$ satisfying $\|M\|_* \leq K$, where $\|\cdot\|_*$ denotes the nuclear norm (the sum of singular values). We

observe Y_{jt}^N for $(j, t) \in \{1, \dots, T\} \times \{1, \dots, J+1\} \setminus \{(1, t) : T_0 + 1 \leq t \leq T\}$. The identifying condition is that $E(u|M) = 0$ and that conditional on M , $\{u_j\}_{j=1}^{J+1}$ is independent across j , where $u_j = (u_{j1}, \dots, u_{jT})' \in \mathbb{R}^T$. The counterfactual proxy is $P_t^N = M_{1t}$ for $1 \leq t \leq T$.

The main challenge is to recover the entire matrix M despite the missing entries $\{Y_{1t}^N : T_0 + 1 \leq t \leq T\}$. The literature on matrix completion considers the model (15) under the assumption of missingness at random and exploits the assumption that the rank of M is low.²⁰ Recently, [Athey et al. \(2018\)](#) introduce this method to study treatment effects in panel data models and point out the unobserved counterfactuals correspond to entries that are missing in a very special pattern, rather than at random. Assuming the usual low rank condition on M , they employ the nuclear norm penalized estimator and provide bounds on the estimation error in the typical setup of causal panel data models.

We take a different approach here since our main goal is hypothesis testing instead of estimation. The key observation is that under the null hypothesis, there are no missing entries in the data. By imposing the null hypothesis, we replace the missing entries with the hypothesized values and obtain a dataset that contains $\{Y_{jt}^N : 1 \leq j \leq J+1, 1 \leq t \leq T\}$. The estimator for M we examine here is closely related to existing nuclear norm regularized estimators and is defined as

$$\hat{M} = \arg \min_{A \in \mathbb{R}^{N \times T}} \sum_{t=1}^T \sum_{j=1}^N (Y_{jt}^N - A_{jt})^2 \quad \text{s.t.} \quad \|A\|_* \leq K, \quad (16)$$

where $K > 0$ is the bound on the nuclear norm of the true matrix. In principle, it can be a sequence that tends to infinity. When M represents a factor structure with strong factors, K can be shown to grow at the rate \sqrt{NT} . Clear guidance on how to choose K is still unavailable, but following [Athey et al. \(2018\)](#) one can use cross-validation.²¹ Alternatively one can use a pilot thresholded SVD estimator to get a sense of what K is, and use a somewhat larger value of K . The following

result guarantees the validity of this estimator in our context under mild regularity conditions.

Lemma 4. *Consider the estimator \hat{M} defined in (16). Assume that $\|M\|_* \leq K$. Let the conditions listed at the beginning of the proof hold. Then, for any $T_0 + 1 \leq t \leq T$, $\hat{P}_t^N - P_t^N = o_p(1)$ and $\frac{1}{T} \sum_{t=K+1}^T (\hat{P}_t^N - P_t^N)^2 = o_p(1)$.*

The result is notable because no sub-Gaussian assumptions are required. The estimator in (16) does not explicitly require a low-rank condition on M . Instead, we impose a growth restriction on K . When M is generated by a strong factor structure and the null hypothesis contains full information on the missing entries, we can choose $K \asymp \sqrt{NT}$ and our consistency result holds as long as $N, T \rightarrow \infty$ and $E(|u_{jt}|^{2+c} | M)$ is uniformly bounded for some $c > 0$. In the case of weak factors, we can choose $K \ll \sqrt{NT}$ and obtain consistency.

4.4 Time Series and Fused Models

As pointed out in Section 2.4, **time series models can be used to model counterfactual proxies with or without control units**. We now discuss low-level conditions under which fitting these models yields estimates good enough for the purpose of our conformal inference approach.

4.4.1 Autoregressive Models

The linear autoregressive model with K lags can be written as

$Y_{it}^N = \rho_0 + \sum_{j=1}^K \rho_j Y_{it-j}^N + u_t$, where $\{u_t\}_{t=1}^T$ is an iid sequence with $E(u_t) = 0$.²² The

counterfactual proxy for Y_{it}^N is $P_t^N = \rho_0 + \sum_{j=1}^K \rho_j Y_{it-j}^N$. We write P_t^N as $P_t^N = y_t' \rho$,

where $y_t = (1, Y_{it-1}^N, Y_{it-2}^N, \dots, Y_{it-K}^N)' \in \mathbb{R}^{K+1}$ and $\rho = (\rho_0, \dots, \rho_K)' \in \mathbb{R}^{K+1}$. The coefficient

vector ρ can be estimated using least squares: $\hat{\rho} = \left(\sum_{t=K+1}^T y_t y_t' \right)^{-1} \left(\sum_{t=K+1}^T y_t Y_{it}^N \right)$. The

estimator for P_t^N is $\hat{P}_t^N = y_t' \hat{\rho}$.

Lemma 5 (Linear AR Model). *Suppose that $\{u_t\}_{t=1}^T$ is an iid sequence with $E(u_1) = 0$ and $E(u_1^4)$ uniformly bounded and the roots of $1 - \sum_{j=1}^K \rho_j L^j = 0$ are uniformly bounded away from the unit circle. Then, for any $T_0 + 1 \leq t \leq T$,*

$$\hat{P}_t^N - P_t^N = o_p(1) \text{ and } \frac{1}{T} \sum_{t=K+1}^T (\hat{P}_t^N - P_t^N)^2 = o_p(1).$$

As mentioned in Section 2.4, we can also apply nonlinear autoregressive models $Y_{1t}^N = \rho(Y_{1t-1}^N, Y_{1t-2}^N, \dots, Y_{1t-K}^N) + u_t$, where ρ is a nonlinear function, in which case the counterfactual proxy is $P_t^N = \rho(Y_{1t-1}^N, Y_{1t-2}^N, \dots, Y_{1t-K}^N)$.

Let $\hat{\rho}$ be an estimator for ρ and $\hat{P}_t^N = \hat{\rho}(Y_{1t-1}^N, Y_{1t-2}^N, \dots, Y_{1t-K}^N)$. This estimator can be parametric, semiparametric, or fully nonparametric and is only required to be consistent.

Lemma 6 (Nonlinear AR Model). *Suppose that (1) $\|\hat{\rho} - \rho\| = O_p(r_T)$ with $r_T = o(1)$ for some appropriate norm $\|\cdot\|$ and*

$$\max_{K+1 \leq t \leq T} |\hat{\rho}(Y_{1t-1}^N, Y_{1t-2}^N, \dots, Y_{1t-K}^N) - \rho(Y_{1t-1}^N, Y_{1t-2}^N, \dots, Y_{1t-K}^N)| \leq \ell_T \|\hat{\rho} - \rho\| \text{ for some}$$

$\ell_T r_T = o(1)$. Then, for any $T_0 + 1 \leq t \leq T$, $\hat{P}_t^N - P_t^N = o_p(1)$ and

$$\frac{1}{T} \sum_{t=K+1}^T (\hat{P}_t^N - P_t^N)^2 = o_p(1).$$

The primitive regularity conditions and the definitions of the neural network estimators possessing these properties can be found, for example, in [Chen and White \(1999\)](#) and [Chen et al. \(2001\)](#).

4.4.2 Fused Panel/Time Series Models with AR Errors

Here we provide generic conditions for the fused panel/time series models described in Section 2.4. In particular, AR models can be used to filter the estimated residuals and obtain near iid errors. In Equation (11) of Section 2.4, we introduce an autoregressive structure in the error terms: $Y_{1t}^N = C_t^N + \varepsilon_t$ and $\varepsilon_t = \rho(\varepsilon_{t-1}) + u_t$, where C_t^N can be specified as a panel data model discussed

before. Due to the autoregressive structure in ε_t , the counterfactual proxy is $P_t^N = C_t^N + \rho(\varepsilon_{t-1})$.

We estimate P_t^N via a two-stage procedure. In the first stage, we estimate C_t^N using the techniques we considered before and obtain say \hat{C}_t^N . In the second stage, we estimate $\rho(\varepsilon_{t-1})$ by fitting an autoregressive model to the estimated residuals $\{\hat{\varepsilon}_t\}_{t=1}^T$, where $\hat{\varepsilon}_t = Y_{1t}^N - \hat{C}_t^N$. For simplicity, we consider a linear model in the second stage estimation. Analogous results can be obtained for more general models. To be specific, assume that $\varepsilon_t = x_t' \rho + u_t$, where $x_t = (\varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-K})' \in \mathbb{R}^K$ and $\rho = (\rho_1, \rho_2, \dots, \rho_K)' \in \mathbb{R}^K$.

Given $\{\hat{\varepsilon}_t\}_{t=1}^T$ from the first-stage estimation, we define $\hat{x}_t = (\hat{\varepsilon}_{t-1}, \hat{\varepsilon}_{t-2}, \dots, \hat{\varepsilon}_{t-K})' \in \mathbb{R}^K$ and $\hat{\rho} = \left(\sum_{t=K+1}^T \hat{x}_t \hat{x}_t' \right)^{-1} \left(\sum_{t=K+1}^T \hat{x}_t \hat{\varepsilon}_t \right)$. To compute the p -value, we use $\{\hat{u}_t\}_{t=K+1}^T$ with $\hat{u}_t = \hat{\varepsilon}_t - \hat{x}_t' \hat{\rho}$ in the permutation. By the following result, this procedure is valid under very mild conditions for the first-stage estimation.

Lemma 7 (AR Errors). *Suppose that $\{u_t\}_{t=1}^T$ is an iid sequence with $E(u_t) = 0$ and $E(u_1^4)$ uniformly bounded and the roots of $1 - \sum_{j=1}^K \rho_j L^j = 0$ are uniformly bounded away from the unit circle. We assume that (1) $\sum_{t=1}^T (\hat{C}_t^N - C_t^N)^2 = o_p(T)$, and (2) $\hat{C}_t^N - C_t^N = o_p(1)$ for $T_0 - K + 1 \leq t \leq T$. Then, for any $T_0 + 1 \leq t \leq T$, $\hat{P}_t^N - P_t^N = o_p(1)$ and $\sum_{t=K+1}^T (\hat{P}_t^N - P_t^N)^2 = o_p(T)$*

Note that the conditions in Lemma 7 for the autoregressive part are the same as in Lemma 5. Consistency of \hat{C}_t^N can be verified using existing results, for example, those in Sections 4.1–4.3.

5 Empirical Application

We revisit the analysis in [Cunningham and Shah \(2018\)](#) who study the impact of decriminalizing indoor prostitution. They consider the case of Rhode Island, where a judge unanticipatedly decriminalized indoor sex work in July 2003 such that, until the recriminalization in November 2009, Rhode Island had decriminalized indoor and prohibited street prostitution.

We focus on the effect of legalizing indoor prostitution on female gonorrhea incidence. Our outcome of interest is log female gonorrhea incidence per 100,000. We use the data on gonorrhea cases from the Center for Disease Control (CDC)'s Gonorrhea Surveillance Program previously analyzed by [Cunningham and Shah \(2018\)](#); see their Section 3 for a detailed description and descriptive statistics. The female gonorrhea series date back to 1985 such that $T_0 = 19$ and $T_* = 6$. Figure 3 displays the raw data for Rhode Island and the rest of the U.S. states.

[Figure 3 around here.]

We apply three different CSC methods: difference-in-differences, canonical SC, and constrained Lasso with $K = 1$. Recall that constrained Lasso nests both difference-in-differences and SC. Following [Cunningham and Shah \(2018\)](#), the set of potential control units includes all other U.S. states and the District of Columbia ($J = 50$). We choose S_1 as our test statistic and report p -values computed based on moving block and iid permutations.²³ All computations were performed in \mathbb{R} ([R Core Team, 2020](#)).

Before turning to the main results, we use the placebo tests proposed in the Appendix to assess the plausibility of the underlying assumptions. Specifically, based on the pre-treatment data, we test $H_0 : \theta_{2003-\tau+1} = \dots = \theta_{2003} = 0$ for $\tau \in \{1, 2, 3\}$. Rejections of this null undermine the credibility of the assumptions underlying our procedure and the inferences on policy effects in the post-treatment period. Table 1 presents the results. Figure 4 complements the formal tests with plots of the residuals from fitting the three models to the pre-treatment data. The placebo

tests and the residual plots provide evidence in favor of the credibility of our inference method in conjunction with SC and, especially, constrained Lasso, but suggest that the difference-in-differences results need to be interpreted with caution.

[Table 1 around here.]

[Figure 4 around here.]

Table 2 reports p -values from testing the null hypothesis of a zero effect:

$$H_0 : \theta_{2004} = \theta_{2005} = \dots = \theta_{2009} = 0. \quad (17)$$

The null hypothesis (17) is rejected at the 10% level based on both permutation schemes and all three methods.

[Table 2 around here.]

Figure 5 displays pointwise 90% confidence intervals. The results are similar for all three methods. While the effect was not or only marginally significant during the first three years, legalizing indoor prostitution significantly decreased the incidence of female gonorrhea thereafter, corroborating the findings by Cunningham and Shah (2018).

[Figure 5 around here.]

To investigate the robustness of our results, we perform a leave-one-out robustness check (e.g., Abadie et al., 2015) to assess whether our findings are driven by a single control state. We iteratively exclude from the control group one of the states for which either the SC or constrained Lasso weights estimated based on the pre-treatment data are non-zero and compute the p -values for testing hypothesis (17). Figure 6 displays the distribution of the resulting p -values. Overall, our results are robust and not driven by a single control state: except for one specification, all results are significant at the 10%-level.

[Figure 6 around here.]

Acknowledgements

We are grateful to Guido Imbens, Jacopo Diquigiovanni, Bruno Ferman, the Co-Editor (Matias Cattaneo), anonymous referees, and many seminar and conference participants for valuable comments. We would like to thank Scott Cunningham and Manisha Shah for sharing the data for the empirical application. Wüthrich is also affiliated with CESifo and the Ifo Institute. Victor Chernozhukov gratefully acknowledges funding by the National Science Foundation. All errors are our own.

Notes

¹We refer to [Doudchenko and Imbens \(2016\)](#), [Gobillon and Magnac \(2016\)](#), and [Abadie \(2019\)](#) for excellent comparative overviews and reviews.

²[Hahn and Shi \(2017\)](#) informally suggest applying a variant of [Andrews \(2003\)](#)'s end-of-sample stability test in the context of SC, and [Ferman and Pinto \(2019a\)](#) use a version of this test in the context of difference-in-differences approaches with few treated groups.

³[Andrews \(2003\)](#) briefly comments on page 1681 (comment 4) that his test can be shown to be asymptotically valid under stationary errors but does not provide a formal result.

⁴In the Appendix, we consider an extension to random policy effects.

⁵In principle, we can relax this assumption by specifying, for example, the scale and quantile shifts in the stochastic shocks that result from the policy, and then working with the resulting model; we leave this extension to future work.

⁶We can also consider other types of permutations; for example, the “iid block” permutations. Specifically, let $\{b_1, \dots, b_K\}$ be a partition of $\{1, \dots, T\}$, then we collect

all the permutations π of these blocks, forming the “iid m-block” permutations Π_{mb} . In our context, choosing $m = T_*$ is natural, though other choices should work as well, similarly to the choice of block size in the time series bootstrap. We refer to CWZ18 for more results on block permutations.

⁷If the estimator of P_t^N is invariant under permutations of the data $\{Z_t\}$ across the time series dimension (which is the case for many estimators in Section 2.3), permuting the residuals $\{\hat{u}_t\}$ is equivalent to permuting the data $\{Z_t\}$.

⁸More generally, other exclusion restrictions and identifying assumptions could be used. See also [Abadie et al. \(2010\)](#), [Ferman and Pinto \(2019b\)](#) and [Ferman \(2019\)](#), who study the behavior of SC when the data are generated by a factor model.

⁹This formulation of canonical SC without covariates is due to [Doudchenko and Imbens \(2016\)](#), who refer to the estimator (4) as “constrained regression”. Note that unlike [Doudchenko and Imbens \(2016\)](#), we estimate w under the null hypothesis based on all the data. We focus on the canonical problem (4) for concreteness. [Abadie et al. \(2010, 2015\)](#) consider a more general version that also includes covariates into the estimation of the weights. Our inference method also works in conjunction with more recently proposed modified versions of SC, such as the augmented SC estimator of [Ben-Michael et al. \(2018\)](#).

¹⁰The idea to relax the non-negativity constraint on the weights is not new. It first appeared in [Hsiao et al. \(2012\)](#), who compared their factor model approach to SC, and also in [Valero \(2015\)](#), who used the cross-validated Lasso to estimate the weights, and in [Doudchenko and Imbens \(2016\)](#), who used cross-validated Elastic Net for estimation of weights. They do not establish the formal properties of these estimators. Here we emphasize another version of relaxing SC, model (5), which leads to constrained Lasso (6). Constrained Lasso demonstrates an excellent theoretical and practical performance: it is tuning-free, performs very well empirically and in simulations, and we prove that it is consistent for

dependent data without any sparsity conditions on the weights and that it satisfies the estimator stability condition required for validity under misspecification. We emphasize that this estimator generally differs from the cross-validated Lasso estimator.

¹¹Factor models are widely used in macroeconomics for causal inference and prediction; see, for example, [Stock and Watson \(2016\)](#) and the references therein. In microeconometrics, factor models are used for estimation of treatment/structural effects; see, for example, [Hansen and Liao \(2019\)](#) who use interactive FE models to estimate the effect of gun prevalence on crime.

¹²Note that [Athey et al. \(2018\)](#)'s analysis applies to a broader collection of problems with general missing data patterns, nesting SC and difference-in-difference problems as special cases.

¹³Note that in our case the sharp null allows us to impute the missing counterfactual response and apply any of the factor estimators to estimate the factor model for the entire data, which is then used for conformal inference. Hence our inference approach does not provide inference for the counterfactual prediction methods given in those papers. Indeed, there, the missing data entries are being predicted using factor models, whereas in our case the missing data entries are known under the null, and we use any form of low-rank approximation or interactive FE model to estimate the model for the entire data under the null hypothesis.

¹⁴We choose to focus on PCA/SVD and the alternating least squares estimator for the following reasons: (1) they are by far the most widely used in practice, (2) the alternating least squares estimator is computationally attractive and easily accommodates unbalanced data.

¹⁵We can also add a moving average component for the errors, but we do not do so for simplicity.

¹⁶We leave the model with persistent feed-forward effects, $Y_{lt}^I = \rho(Y_{l(t-1)}^I) + \theta_t + u_t$, to future work.

¹⁷While our general results in this section are non-asymptotic, some of the analysis in Section 4 will not be non-asymptotic in nature.

¹⁸To simplify the exposition, we do not include an intercept in Lemma 1. Similar arguments could be used to prove an analogous result with an unconstrained intercept.

¹⁹Note that PCA amounts to singular value decomposition, which can be computed using polynomial time algorithms, (e.g., [Trefethen and Bau III, 1997](#), Lecture 31).

²⁰See, for example, [Candès and Recht \(2009\)](#), [Recht et al. \(2010\)](#), [Candès and Plan \(2011\)](#), [Koltchinskii et al. \(2011\)](#), [Negahban et al. \(2011\)](#), [Rohde and Tsybakov \(2011\)](#), and [Chatterjee \(2015\)](#).

²¹The properties of cross-validation remain unknown in these settings.

²²Here the model seems different, but Section 2.4's model implies this one with

$$\rho_0 = \mu(1 - \sum_{j=1}^K \rho_j).$$

²³To keep computation tractable, we randomly sample 10,000 iid permutations with replacement.

References

Abadie, A. (2019). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *forthcoming at the Journal of Economic Literature*.

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. (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., Shah, D., and Shen, D. (2018). Robust synthetic control. *The Journal of Machine Learning Research*, 19(1):802–852.

Andrews, D. W. (2003). End-of-sample instability tests. *Econometrica*, 71(6):1661–1694.

Antoch, J. and Huskova, M. (2001). Permutation tests in change point analysis. *Statistics & Probability Letters*, 53(1):37 – 46.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2018). Synthetic difference in differences. *arXiv:1812.09970*.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2018). Matrix completion methods for causal panel data models. Working Paper 25132, National Bureau of Economic Research.

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.

Ben-Michael, E., Feller, A., and Rothstein, J. (2018). The augmented synthetic control method. *arXiv:1811.04170*.

Brockwell, P. J. and Davis, R. A. (2013). *Time series: theory and methods*. Springer Science & Business Media.

Candès, E. J. and Plan, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772.

Carrasco, M. and Chen, X. (2002). Mixing and moment properties of various garch and stochastic volatility models. *Econometric Theory*, 18(1):17–39.

Carvalho, C., Masini, R., and Medeiros, M. C. (2018). Arco: An artificial counterfactual approach for high-dimensional panel time-series data. *Journal of Econometrics*, 207(2):352–380.

Cattaneo, M. D., Feng, Y., and Titiunik, R. (2021). Prediction intervals for synthetic control methods. *arXiv:1912.07120*.

Chan, M. and Kwok, S. (2016). Policy evaluation with interactive fixed effects. The University of Sidney, Economics Working Paper Series, 2016–11.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214.

Chen, X., Racine, J., and Swanson, N. R. (2001). Semiparametric arx neural-network models with an application to forecasting inflation. *IEEE Transactions on neural networks*, 12(4):674–683.

Chen, X. and White, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691.

Chernozhukov, V., Hansen, C., and Liao, Y. (2017). A lava attack on the recovery of sums of dense and sparse signals. *The Annals of Statistics*, 45(1):39–76.

Chernozhukov, V., Wüthrich, K., and Yinchu, Z. (2018). Exact and robust conformal inference methods for predictive machine learning with dependent data. In Bubeck, S., Perchet, V., and Rigollet, P., editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 732–749. PMLR.

Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2019). Practical and robust t-test based inference for synthetic control and related methods. arXiv:1812.10820.

Conley, T. G. and Taber, C. R. (2011). Inference with "difference in differences" with a small number of policy changes. *The Review of Economics and Statistics*, 93(1):113–125.

Cunningham, S. and Shah, M. (2018). Decriminalizing indoor prostitution: Implications for sexual violence and public health. *The Review of Economic Studies*, 85(3):1683–1715.

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.

Dufour, J.-M., Ghysels, E., and Hall, A. (1994). Generalized predictive tests and structural change analysis in econometrics. *International Economic Review*, 35(1):199–229.

Ferman, B. (2019). On the properties of the synthetic control estimator with many periods and many controls. *arXiv:1906.06665*.

Ferman, B. and Pinto, C. (2019a). Inference in differences-in-differences with few treated groups and heteroskedasticity. *The Review of Economics and Statistics*, 101(3):452–467.

Ferman, B. and Pinto, C. (2019b). Synthetic controls with imperfect pre-treatment fit. *arXiv:1911.08521*.

Firpo, S. and Possebom, V. (2018). Synthetic control method: Inference, sensitivity analysis and confidence sets. *Journal of Causal Inference*, 6(2).

Fisher, R. (1935). *The Design of Experiments*. Oliver & Boyd.

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. (2017). Synthetic control and inference. *Econometrics*, 5(4):1–12.

Hamilton, J. D. (1994). *Time series analysis*. Princeton: Princeton University Press.

Hansen, C. and Liao, Y. (2019). The factor-lasso and k-step bootstrap approach for inference in high-dimensional economic applications. *Econometric Theory*, 35(3):465–509.

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.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Science & Business Media.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lei, J., Robins, J., and Wasserman, L. (2013). Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287.

Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96.

Li, K. (2018). Inference for factor model based average treatment effects. *Available at SSRN 3112775*.

Li, K. T. (2020). Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association*, 115(532):2068–2083.

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.

Negahban, S., Wainwright, M. J., et al. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097.

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.

Politis, D. N. (2015). *Model-free prediction and regression: a transformation-based approach to inference*. Springer, New York.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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.

Recht, B., Fazel, M., and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501.

Rohde, A. and Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930.

Romano, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85(411):686–692.

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

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172.

Shaikh, A. M. and Toulis, P. (2019). Randomization tests in observational studies with staggered adoption of treatment. arXiv:1912.10610.

Stock, J. and Watson, M. (2016). Chapter 8 - dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. volume 2 of *Handbook of Macroeconomics*, pages 415–525. Elsevier.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288.

Trefethen, L. N. and Bau III, D. (1997). *Numerical linear algebra*, volume 50. Siam.

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

Vovk, V., Gammernan, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.

Vovk, V., Nouretdinov, I., and Gammernan, A. (2009). On-line predictive linear regression. *The Annals of Statistics*, 37(3):1566–1590.

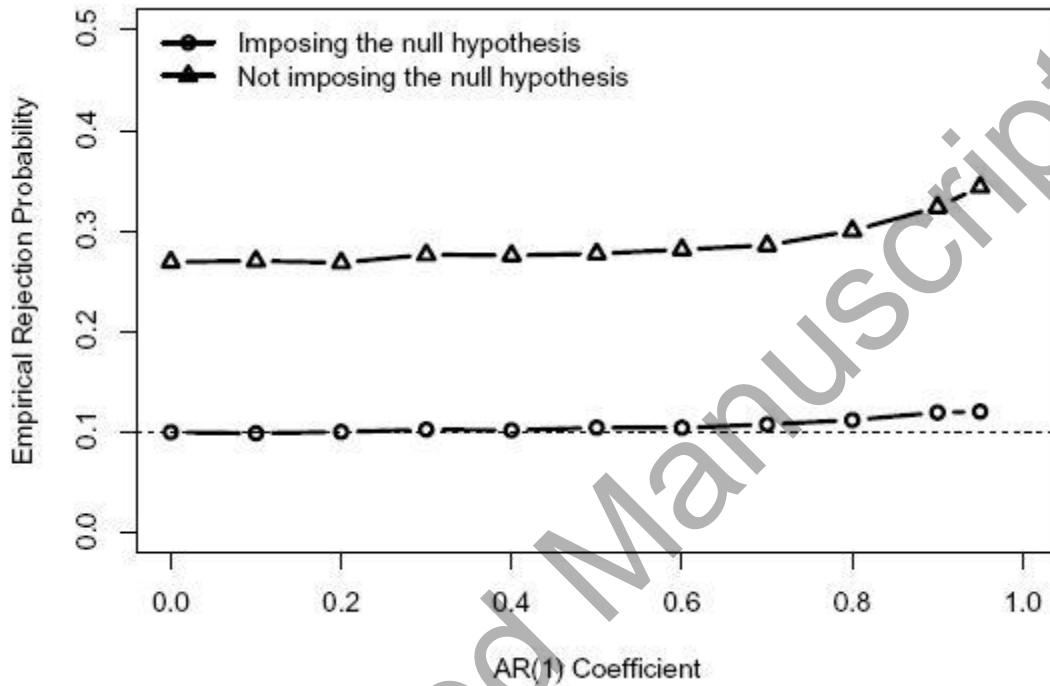
White, H. (1996). *Estimation, inference and specification analysis*. Number 22. Cambridge University Press.

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

Zeileis, A. and Hothorn, T. (2013). A toolbox of permutation tests for structural change. *Statistical Papers*, 54(1):931–954.

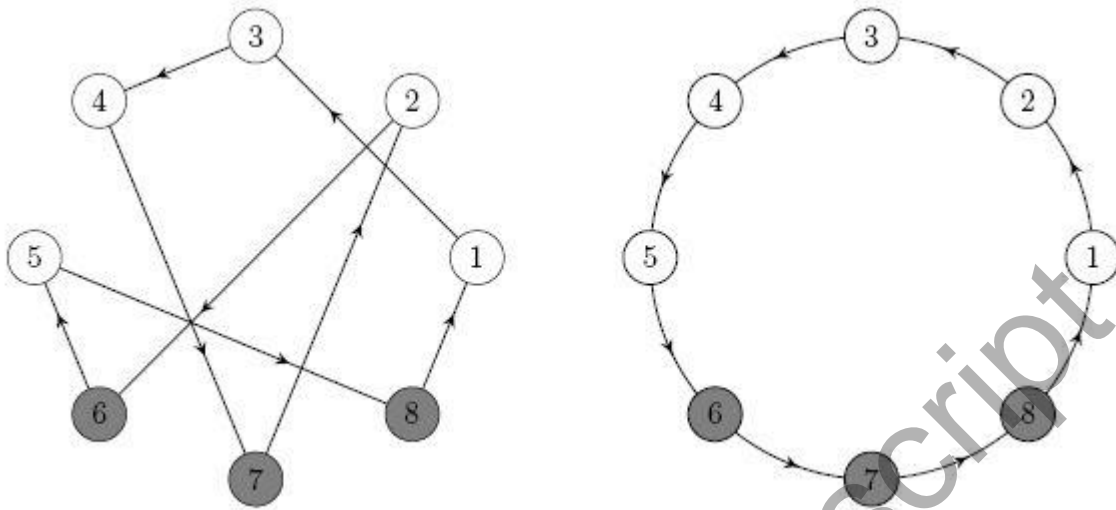
Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Fig. 1 Small Sample Size Properties (Nominal Level: 10%)



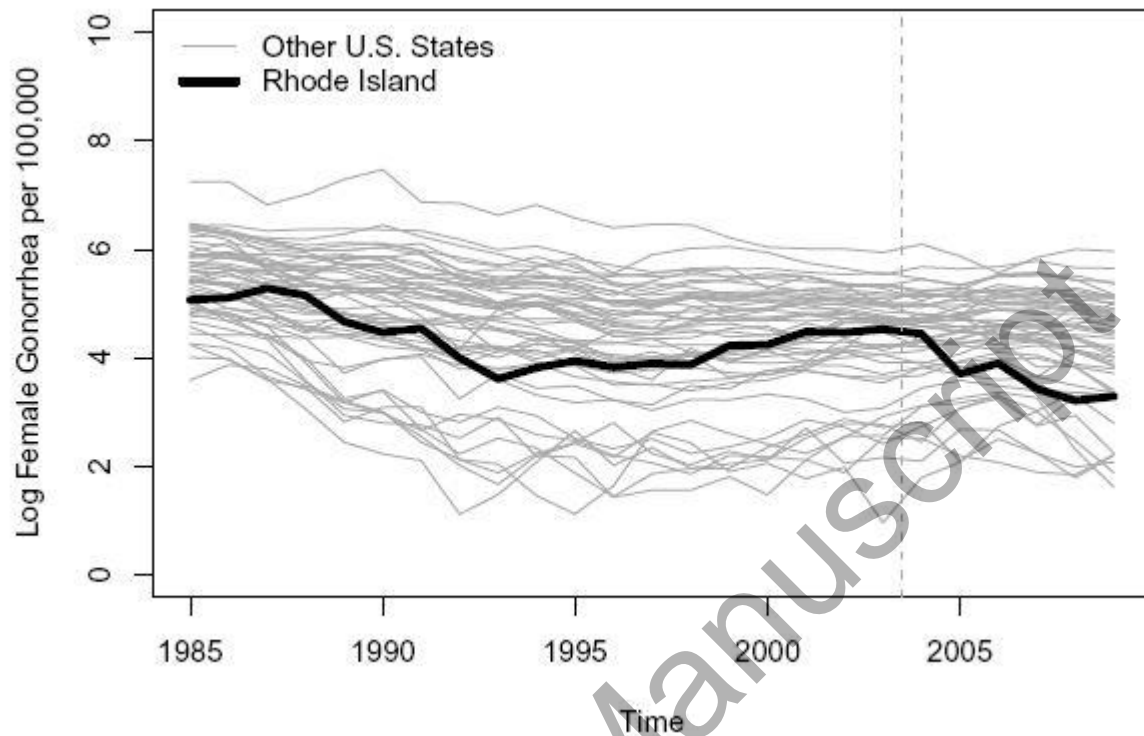
Notes: Empirical rejection probability from testing $H_0 : \theta_{T_0+1} = 0$. The data are generated as $Y_{1t}^N = \sum_{j=2}^{J+1} w_j Y_{jt}^N + u_t$, where $Y_{jt}^N \sim N(0,1)$ is iid across (j, t) , $\{u_t\}$ is a Gaussian AR(1) process, $(w_2, \dots, w_{J+1})' = (1/3, 1/3, 1/3, 0, \dots, 0)'$, $T_0 = 19$, and $J = 50$. The weights are estimated using the canonical SC method (cf. Section 2.3.2).

Fig. 2 Graphical Illustration Permutations



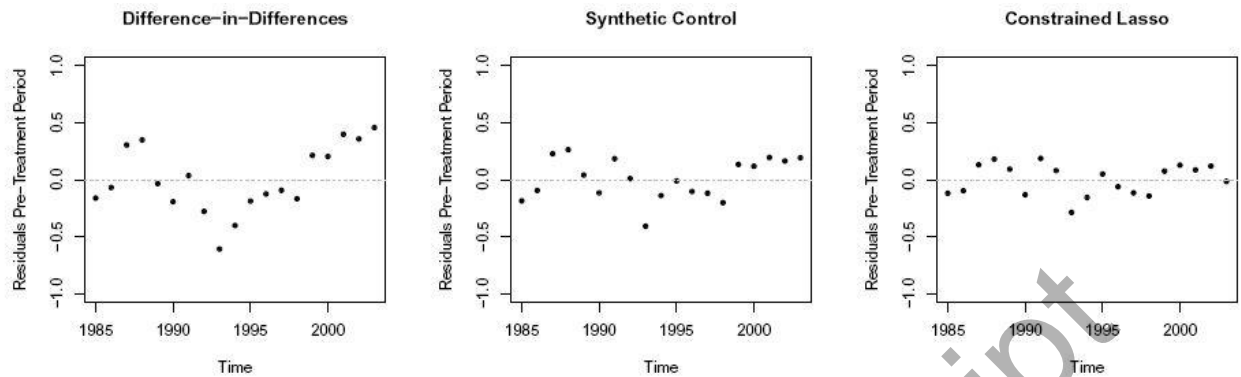
Notes: The left figure gives an example of an iid permutation of $\{1, 2, 3, 4, 5, 6, 7, 8\}$. The right figure gives an example of a moving block permutation of $\{1, 2, 3, 4, 5, 6, 7, 8\}$. $T_0 = 5, T_* = 3$. Pre-treatment periods are white; post-treatment periods are gray.

Fig. 3 Raw Data



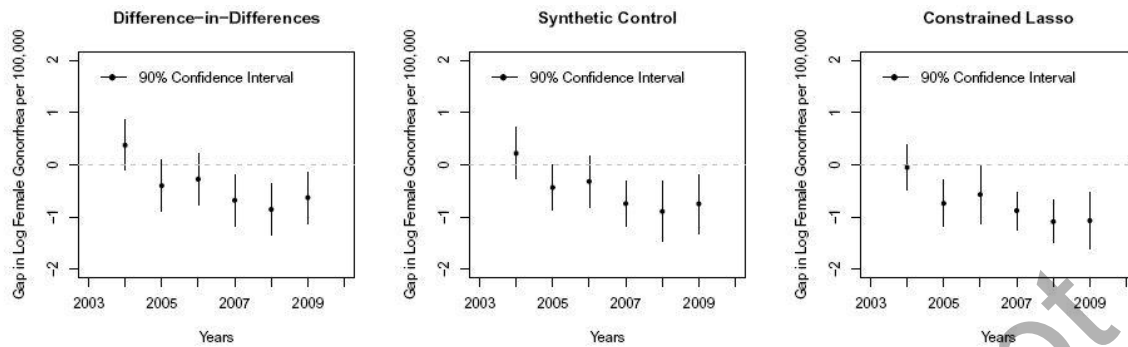
Notes: Data are from Cunningham and Shah (2018). The figure shows the raw state-level data on log female gonorrhea cases per 100,000.

Fig. 4 Graphical Placebo Checks



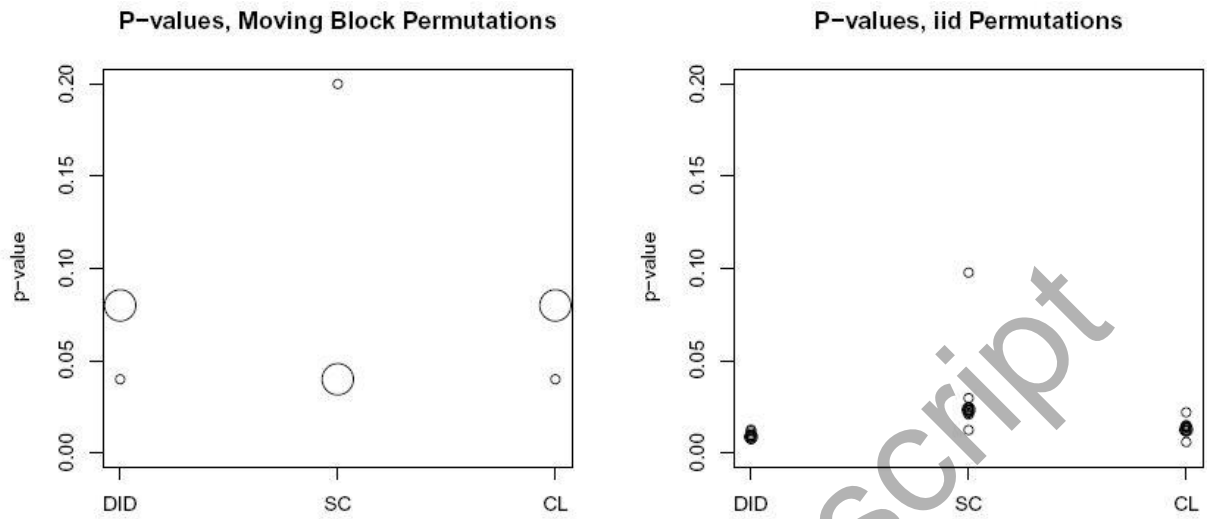
Notes: Data are from [Cunningham and Shah \(2018\)](#). The figure plots the pre-treatment residuals estimated using difference-in-differences, SC, and constrained Lasso.

Fig. 5 Pointwise Confidence Intervals



Notes: Data are from [Cunningham and Shah \(2018\)](#). The figure plots pointwise 90% confidence intervals computed using Algorithm 1.

Fig. 6 Leave-one-out Robustness Checks



Notes: Data are from [Cunningham and Shah \(2018\)](#). This figure shows the distribution of p -values from testing null hypothesis (17), leaving-out one of the control states with non-zero weight at the time. The size of the circles is proportional to the number of p -values. DID: difference-in-differences; SC: synthetic control; CL: constrained Lasso.

Table 1 Placebo Specification Tests

	Moving Block Permutations			iid Permutations		
τ	Diff-in-Diffs	Synth. Control	Constr. Lasso	Diff-in-Diffs	Synth. Control	Constr. Lasso
1	0.11	0.32	1.00	0.11	0.31	1.00
2	0.16	0.32	0.89	0.06	0.31	0.93
3	0.11	0.26	1.00	0.03	0.25	0.95
.						

Notes: Data are from [Cunningham and Shah \(2018\)](#). Table shows p -values from testing $H_0 : \theta_{2003-\tau+1} = \dots = \theta_{2003} = 0$ for $\tau \in \{1, 2, 3\}$ based on the pre-treatment data

Table 2 Zero Effect Null Hypothesis

Moving Block Permutations			iid Permutations		
Diff-in-Diffs	Synth. Control	Constr. Lasso	Diff-in-Diffs	Synth. Control	Constr. Lasso
0.08	0.04	0.08	0.01	0.03	0.01

Notes: Data are from Cunningham and Shah (2018). Table shows p -values from testing $H_0 : \theta_{2004} = \theta_{2005} = \dots = \theta_{2009} = 0$.