An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls [*]

Victor Chernozhukov[†] Kaspar Wüthrich[‡] Yinchu Zhu[§]

December 6, 2018

Abstract

We introduce new inference procedures for counterfactual and synthetic control methods for evaluating policy effects. Our methods work in conjunction with many different approaches for estimating the counterfactual mean outcome in the absence of a policy intervention. Examples include difference-in-differences, synthetic controls, factor and matrix completion models, and (fused) time series panel data models. Our procedures have a double justification. (i) If the residuals from estimating the counterfactuals are exchangeable, our procedures achieve finite sample size control without any assumptions on the specific approach used to estimate the counterfactuals. (ii) If the data exhibit dynamics and dependence, our procedures are approximately valid under weak conditions on the method used to estimate the counterfactuals. We verify these conditions for representative methods from each group listed above. Our approach demonstrates an excellent performance in simulation experiments, and is taken to a data application, where we re-evaluate the consequences of decriminalizing indoor prostitution.

Keywords: policy evaluation, panel data, conformal inference, permutation inference, model-free validity, difference-in-differences, synthetic control, factor model, interactive fixed effects model, matrix completion, constrained Lasso, machine learning

[*] We are grateful to Guido Imbens and seminar participations at LMU Munich, Penn State, Princeton ORFE, the University of Chicago, the University of Southern California, the University of Wisconsin Madison, UC Berkeley, UC Los Angeles, UC San Diego, Vanderbilt and University of Connectitcut, and conference participants at Workshop Machine Learning and Causal Inference at Erasmus University Rotterdam, the New Frontier in Econometrics Conference at University of Connecticut, the CEME Conference 2018, and the California Econometrics Conference 2018 for valuable comments. We would like to thank Scott Cunningham and Manisha Shah for sharing the data for the empirical application. All errors are our own.

[†] Massachusetts Institute of Technology; 50 Memorial Drive, E52-361B, Cambridge, MA 02142, USA; Email: vchern@mit.edu

[‡] UC San Diego; 9500 Gilman Dr., La Jolla, CA 92093, USA; Email: kwuthrich@ucsd.edu

[§] University of Oregon; 1208 University St, Eugene, OR 97403, USA; Email: yzhu6@uoregon.edu

1 Introduction

We consider the problem of making inference 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 a number of periods before and 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 economic research and there are various different approaches for estimating the policy effects of interest. A non-exhaustive list of examples includes difference-in-differences methods (e.g., Ashenfelter and Card, 1985; Card and Krueger, 1994; Bertrand et al., 2004; Athey and Imbens, 2006; Angrist and Pischke, 2008), synthetic control approaches (e.g., Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015; Li, 2017), penalized regression models for synthetic control (e.g., Valero, 2015; Doudchenko and Imbens, 2016; Li and Bell, 2017; Carvalho et al., 2017), factor, matrix completion, and interactive fixed effects models (e.g., Bai, 2003; Pesaran, 2006; Bai, 2009; Hsiao et al., 2012; Kim and Oka, 2014; Gobillon and Magnac, 2016; Chan and Kwok, 2016; Xu, 2017; Athey et al., 2017; Amjad et al., 2017; Li, 2018), matching methods (e.g., Heckman et al., 1997, 1998; Dehejia and Wahba, 2002), as well as standard time series models. Doudchenko and Imbens (2016) and Gobillon and Magnac (2016) provide comparative overviews. We refer to these methods as counterfactual and synthetic control (CSC) methods.

The main objective and contribution of this paper is to provide inference procedures for policy effects estimated by CSC methods. There are several practical issues which render inference in CSC settings challenging. First, the data typically exhibit dynamics and serial dependence. Second, because there is only one treated unit and the number of post-treatment periods, T_*, is typically small, treatment effects cannot be consistently estimated, even if the number of pre-treatment periods, T_0, is very large. Third, the number of (potential) control units, J, is often of the same order as T_0, which leads to a need for some regularization. This paper develops an inference approach that addresses all these challenges. Since the applicability and credibility of CSC methods typically requires a sizable number of pre-treatment periods, as pointed out by for instance by Abadie et al. (2015), we shall focus on inference in this setting.[1]

We analyze a general counterfactual modeling framework (CMF) that nests and gen-

[1] Referring to the synthetic control method, one of the leading examples analyzed in this paper, Abadie et al. (2015) write: “The applicability of the method requires a sizable number of preintervention periods. The reason is that the credibility of a synthetic control depends upon how well it tracks the treated unit's characteristics and outcomes over an extended period of time prior to the treatment. We do not recommend using this method when the pretreatment fit is poor or the number of pretreatment periods is small” (p. 500).

eralizes many traditional and new methods for counterfactual analysis. Specifically, we focus on models which are able to generate mean-unbiased proxies, P_t^N, for the counterfactual outcome of the treated unit in the absence of the policy intervention, Y_{1t}^N:

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

The policy effect in period t is given by $\alpha_t = Y_{1t}^I - Y_{1t}^N$, where Y_{1t}^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-intervention period: $\alpha = \{\alpha_t\}_{t=T_0+1}^{T_0+T_*}$. Specifically, we postulate a null trajectory, $\alpha^o = \{\alpha_t^o\}_{t=T_0}^{T_0+T_*}$, and test the sharp null hypothesis $H_0 : \alpha = \alpha^o$. We also consider the problem of testing hypotheses about per-period treatment effects, $H_0 : \alpha_t = \alpha_t^o$, and propose a simple algorithm for constructing confidence intervals for α_t via test inversion.

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 hypothesis $H_0 : \alpha_{T_0+1} = \alpha_{T_0+1}^o$, we can compute Y_{1t}^N and, thus, $u_t = Y_{1t}^N - P_t^N$ for all time periods. If the stochastic shock process, $\{u_t\}$, is stationary and weakly dependent, the distribution of the stochastic shock in the post-treatment period, u_{T_0+1}, should be the same as the distribution of the shocks in the pre-treatment period, $\{u_1, \ldots, u_{T_0}\}$. We operationalize this idea by proposing conformal/permutation inference procedures in which p-values are obtained by permuting the estimated residuals, $\{\hat{u}_t\}$, across the time series dimension. The proposed approach has a double justification:[2]

(i) Model-Free Exact Validity under Strong Assumptions on the Data.

If the estimated residuals, $\{\hat{u}_t\}$, are exchangeable, our inference procedures achieve finite-sample (non-asymptotic) size control without any assumptions on the method used to estimate the counterfactual mean proxy, P_t^N. As a consequence, our method controls size under arbitrary misspecification and is fully robust against overfitting. Exchangeability of $\{\hat{u}_t\}$ is implied, for example, if the data are i.i.d. across time, but holds more generally.

(ii) Approximate Validity under Weak Assumptions on the Data and the Estimators.

Our procedures achieve approximate finite-sample size control under two different sets of conditions.

[2] Our title is inspired by Chung and Romano (2013), who show that permutation tests have a double justification under two different sets of assumptions.

(a) If the data exhibit dynamics and serial dependence, but $\{u_t\}$ is stationary and weakly dependent, our procedures are approximately valid if the estimator of the counterfactual mean proxy, P_t^N, is consistent (pointwise and in the prediction norm). Consistency can be verified for many different CSC methods. We provide concrete sets of sufficient conditions for a representative selection of methods, including canonical synthetic controls, factor, matrix completion and interactive fixed effects models, linear and nonlinear time series models, as well as fused time series panel data models.

(b) If the model for the counterfactual mean proxy, P_t^N, is misspecified and the estimator of P_t^N is inconsistent, our procedures are approximately valid, provided that the data are stationary and weakly dependent and that the estimator of P_t^Nsatisfies a certain *stability* condition. This condition requires that the estimator is stable under perturbations in a few observations. It is implied, for instance, if the estimator of P_t^N is consistent for a "pseudo-true" parameter value, but is shown to hold much more generally.

We would like to emphasize two additional contributions of this paper that may be of substantial independent interest. First, we introduce the ℓ_1-constrained least squares estimator (e.g., Raskutti et al., 2011), which we will refer to as *constrained Lasso*, as an essentially tuning-free alternative to existing penalized regression estimators in settings with potentially very many control units. Constrained Lasso nests both canonical synthetic control and difference-in-differences as special cases and thus provides a unifying approach to the regression-based estimation of counterfactual mean proxies. We establish pointwise consistency and consistency in the prediction norm of the constrained Lasso estimator of P_t^N in non-Gaussian settings with dependent data. Importantly, these results do not rely on any sparsity assumptions. In addition, we provide a set of sufficient conditions under which constrained Lasso satisfies the estimator stability condition required for the approximate validity of our method under misspecification. Our conditions do not require the constrained Lasso estimator to converge to anything nor do they rely on sparsity of the best linear projection as imposed for instance by Bühlmann and van de Geer (2015). Second, as a byproduct of our theoretical analysis of constrained Lasso, we obtain new theoretical consistency results for synthetic control estimators in settings with potentially very many control units.

We develop three extensions of our main results. First, we demonstrate that our method can be modified to test hypotheses about average effects over time. Second, we show how to make inference with multiple treated units. Finally, we develop easy-to-implement specification tests for assessing the plausibility of the key assumptions underlying our procedures.

Our approach demonstrates an excellent performance in simulation experiments, and is taken to a data application, where we re-visit the analysis of the causal effect of decriminalizing indoor prostitution on rape rates and sexually transmitted infections by Cunningham and Shah (2018).

1.1 Related Literature

Conceptually, our procedure builds 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, more broadly, 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 also Rubin (1984) for a Bayesian justification. On a more general conceptual level, our approach is also connected to transformation-based approaches to model-free prediction (e.g., Politis, 2015).

Conformal inference, a form of permutation inference, is a distribution-free approach for forming prediction intervals. The basic idea is classical. Let $\{Y_1, \ldots, Y_T\}$ be a random sample drawn from a distribution P. To decide whether a new draw $Y_{T+1} = y$ should be included in the prediction set, we test the hypothesis that $Y_{T+1} = y$. A distribution-free and valid p-value can be constructed based on the quantile of the empirical distribution of the augmented sample $\{Y_1, \ldots, Y_T, y\}$. There are two important conceptual differences between our approach and the literature on conformal prediction. First, our goal is to make inference on the causal effects of policy interventions, whereas the goal of the existing conformal prediction methods is to construct distribution-free prediction intervals. Second, since we study an aggregate times series setting, we have to deal with dynamics and general patterns of data dependence, whereas the distribution-free validity of conformal prediction crucially relies on the exchangeability of the data; see Chernozhukov et al. (2018) for an extension to weakly dependent data.

The proposed inference procedures are further related to Andrews (2003)'s end-of-sample stability test based on subsampling. Besides a different focus (inference on policy effects vs. testing for structural breaks), there are several major differences. First, our procedures are exactly valid under exchangeability and we obtain approximate finite sample bounds under weak conditions on the estimators, while such properties have not been established for Andrews (2003)'s test. Second, we demonstrate that our methods are valid under misspecification, provided that the estimator satisfies a certain stability condition, whereas Andrews (2003) assumes correct specification. A third important difference is that some of our results only require stationarity and weak dependence of the stochastic process, $\{u_t\}$, while Andrews (2003)'s test is based on stationarity of the data.[3] A fourth major difference is that our procedures work in conjunction with many modern high-dimensional estimators, whereas Andrews (2003) focuses on low-dimensional GMM-type models. Hahn and Shi (2016, Section 5) informally suggest applying the end-of-sample stability test in the context of synthetic control methods and Ferman and Pinto (2017a) use a version of this test in the context of difference-in-differences approaches with few treated groups.

Our paper contributes to the literature on inference for CSC methods with few treated units. One part of the literature considers a finite-population approach, which relies on the assumption that potential outcomes are fixed but a priori unknown and that, conditional on observables, the treatment assignment is random (Firpo and Possebom, 2017). These assumptions justify the application of permutation tests similar to Fisher (1935)'s randomization test. For instance, Abadie et al. (2010, 2015) permute which unit is assigned to the treatment and then compare the actual treatment effect estimates to the permutation distribution.[4] We refer to Firpo and Possebom (2017) and Ferman and Pinto (2017b) for excellent discussions of the theoretical aspects of these testing procedures. While finite-population permutation approaches have traditionally been employed in conjunction with synthetic control methods, similar ideas be applied to a much broader class of methods, including difference-in-differences approaches, elastic net, and best subset selection; see Doudchenko and Imbens (2016). Our approach will instead carry out the permutations over stochastic errors in the potential outcomes with respect to time, and not the cross-sectional units. These types of permutations rely on weak dependence of stochastic errors over time rather than exchangeability across treated units.

Another strand of the literature considers asymptotic inference for CSC models. Asymptotic approaches often focus on testing hypotheses about average effects over time and require the number of pre-period and post-treatment periods to tend to infinity. Li and Bell (2017), Carvalho et al. (2017), and Li (2017) introduce inference methods based on penalized and constrained regression methods. Inference procedures for policy effects estimated based on factor and interactive fixed effects models are proposed by Hsiao et al. (2012), Gobillon and Magnac (2016), Chan and Kwok (2016), Li and Bell (2017), Xu (2017),

[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.

[4] Conley and Taber (2011) propose a conceptually related inference procedure for difference-in-differences models with few policy changes, which exploits cross-sectional information about the distribution of the unobserved components.

and Li (2018). By contrast, our approach will instead be based on sharp null hypotheses and permutation distributions, and will be shown to be exactly valid under strong exchangeability assumptions and approximately valid under stationarity and weak dependence assumptions as well as mild conditions on the estimators of the counterfactual mean proxies. We verify these conditions for many different methods including constrained least squares estimators, factor models, and interactive fixed effects estimators.

1.2 Plan of the Paper

We introduce some frequently used notations. For $q \geq 1$, the ℓ_q-norm of a vector is denoted by $\|\cdot\|_q$. We use $\|\cdot\|_0$ to denote the number of nonzero entries of a vector; $\|\cdot\|_\infty$ is used to denote the maximal absolute value of entries of a vector. We also use the notation $a \lesssim b$ to denote $a \leq cb$ for some constant $c > 0$ that does not depend on the sample size. We also use the notation $a \asymp b$ to denote $a \lesssim b$ and $b \lesssim a$. 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.

The remainder of this paper is structured as follows. Section 2 introduces our basic modeling framework, the proposed inference method, and various models for the counterfactual mean proxies. In Section 3 we establish the finite sample validity of our procedure if the residuals are exchangeable and the approximate uniform validity with dependent data. Section 4 discusses extensions. In Sections 5 and 6 we verify the conditions for the approximate validity of our method for several representative CSC estimators. Section 7 presents some simulation evidence on the finite sample properties of our approach. In Section 8 we re-evaluate the causal effect of decriminalizing indoor prostitution on rape rates and sexually transmitted infections. Section 9 concludes. All proofs are collected in the appendix.

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 a policy. Extensions to more than one treated unit are discussed in Section 4.2. Our typical setting is where T_* is short compared to T_0. There may be other units which are not exposed to the treatment, and they will be introduced below. We denote the observed outcome of the treated unit by Y_{1t}. Our analysis is developed within the potential (latent) outcome framework (Neyman, 1923; Rubin, 1974). Potential outcomes with and without the policy are denoted as Y_{1t}^I and Y_{1t}^N, respectively. The policy effect in period t is given by $\alpha_t = Y_{1t}^I - Y_{1t}^N$.

Our conformal inference method will rely on the following counterfactual modeling framework:

Assumption 1 (Counterfactual Model). *Let $\{P_t^N\}$ be a given sequence of mean unbiased signals or proxies for the counterfactual outcomes $\{Y_{1t}^N\}$ in the absence of the policy intervention, that is $\{E\left(P_t^N\right)\} = \{E\left(Y_{1t}^N\right)\}$. Let $\{\alpha_t\}$ be a fixed treatment effect sequence such that $\alpha_t = 0$ for $t \leq T_0$, so that potential outcomes under the intervention are given by $\{Y_{1t}^I\} = \{Y_{1t}^N + \alpha_t\}$. In other words, the following system of structural equations holds:*

$$\left.\begin{aligned} Y_{1t}^N &= P_t^N + u_t \\ Y_{1t}^I &= P_t^N + \alpha_t + u_t \end{aligned}\right| \quad E(u_t) = 0, \quad t = 1, \ldots, T, \tag{CMF}$$

where $\{u_t\}$ is a centered stationary stochastic process. Observed outcomes are related to potential outcomes as

$$Y_{1t} = Y_{1t}^N + D_t\left(Y_{1t}^I - Y_{1t}^N\right), \quad t = 1, \ldots, T,$$

where $D_t = 1\left(t > T_0\right)$ is the treatment indicator.

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

$$E\left(P_t^N\right) = E\left(Y_{1t}^N\right).$$

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 i.i.d. or more generally a stationary and weakly dependent process. In principle, the treatment effect sequence $\{\alpha_t\}$ can be allowed to be random, and we can interpret our model and the results as holding conditional on a given $\{\alpha_t\}$. Hence, there is not much loss of generality in assuming that the sequence is fixed. Assumption 1 also postulates that the stochastic shock sequence will be invariant under the intervention. This is the key identifying assumption. In 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. The CMF nests many traditional and new methods for counterfactual policy analysis, including difference-in-differences methods, canonical synthetic control, constrained and penalized regressions for synthetic control, factor/matrix completion models for panel data, interactive fixed effects panel models, univariate time series models, as well as fused time series panel data models.

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 $j = 2, \ldots, J + 1$. 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, Y_{jt}^N, so that

$$Y_{jt} = Y_{jt}^N, \quad j = 2, \ldots, J + 1, \quad t = 1, \ldots, T.$$

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.

In a nutshell, our inference approach will postulate a null trajectory:

$$\alpha^o = \{\alpha_t^o\}_{t=T_0}^T.$$

Under Assumption 1, we can subtract α_t^o from the observed Y_{1t} in post-treatment period to obtain Y_{1t}^N. Using appropriate panel data or time series approaches, we can model, identify, and estimate P_t^N to back out the distribution of $\{u_t\}$ under the null hypothesis. We will use this distribution to compute the null distribution of the relevant test statistic, and then compare the actual observed statistic against this distribution. We will justify this procedure as exactly valid under strong assumptions, and approximately valid under weak assumptions.

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, $\alpha = (\alpha_{T_0+1}, \ldots, \alpha_T)'$. Our main hypothesis of interest is

$$H_0 : \alpha = \alpha^o, \tag{1}$$

where $\alpha^o = \left(\alpha_{T_0+1}^o, \ldots, \alpha_T^o\right)'$ 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 treatment in the post treatment period since $Y_{1t}^N = Y_{1t}^I - \alpha_t = Y_{1t} - \alpha_t$. Our procedure can be extended to test hypotheses about average effects over time as discussed in Section 4.1.

To describe our procedure, we write the data under the null as $\mathbf{Z} = (Z_1, \ldots, Z_T)'$, where

$$Z_t = \begin{cases} \left(Y_{1t}^N, Y_{2t}^N, \ldots, Y_{J+1t}^N, X_{1t}', \ldots, X_{J+1t}'\right)', & t \leq T_0 \\ \left(Y_{1t} - \alpha_t^o, Y_{2t}^N, \ldots, Y_{J+1t}^N, X_{1t}', \ldots, X_{J+1t}'\right)', & t > T_0. \end{cases}$$

Using one of the methods described below, we will obtain a counterfactual proxy estimate, $\hat{P}_t^N$, using $\mathbf{Z}$, and obtain the residuals

$$\hat{u} = (\hat{u}_1, \ldots, \hat{u}_T)', \quad \hat{u}_t = Y_{1t}^N - \hat{P}_t^N, \quad t = 1, ..., T.$$

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

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

In applications we will mostly be using S_1 by setting $q = 1$, which behaves well under heavy-tailed data. We note that other test statistics could be considered as well. When the nature of the statistic is not essential, we write $S = S_q$. Note that S is constructed such that high values indicate rejection.

Remark 1. When capturing deviations in average treatment effect $T_*^{-1} \sum_{t=T_0+1}^{T} \alpha_t$ it is useful to consider the statistic of the form:

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

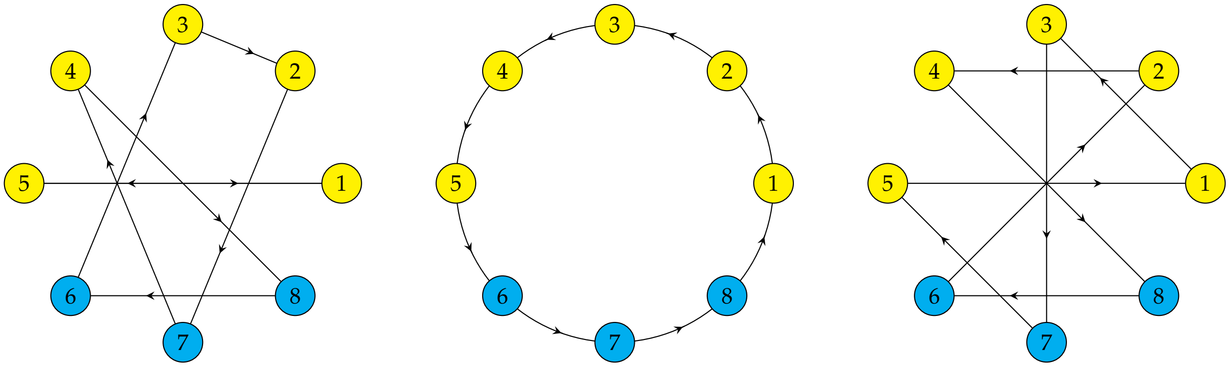
We use permutations to compute p-values. A permutation π is a one-to-one mapping $\pi : \{1, \ldots, T\} \mapsto \{1, \ldots, T\}$. We denote the set of all permutations under study as Π. Throughout the paper we assume that Π contains the identity map $\mathbb{I}$. We mainly focus on two different sets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

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

The choice of Π does not matter for the exact finite sample validity of our procedures if the residuals are exchangeable. However, the set of all i.i.d. permutations will typically have more elements than the set of moving block permutations. For the approximate finite sample results under estimator consistency, the choice of Π depends on the the assumptions that we are willing to impose on the stochastic shock sequence $\{u_t\}$. If $\{u_t\}$ is i.i.d., approximate size control can be established based on both sets of permutations. On the other hand, if $\{u_t\}$ exhibits serial dependence, we will have to rely on moving block permutations.

Remark 2. We can also consider the "i.i.d. block" permutations. We divide the data up into non-overlapping $K = T/m$ blocks of size m. Then we construct the "i.i.d" permutations of all blocks. Specifically, let $\{b_1, \ldots, b_K\}$ be a partition of $\{1, \ldots, T\}$, then we collect all the permutations π of these blocks, forming the "i.i.d. m-block" permutation 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.

Figure 1: Permutations: "I.I.D", "Moving Blocks", "I.I.D. Blocks".



Notes: The left figure gives an example of an "i.i.d" permutation, the middle figure gives the "moving block" permutation, the right figure gives an "i.i.d. block" permutation. In the "i.i.d" permutation, $\pi : \{1, 2, 3, 4, 5, 6, 7, 8\} \mapsto \{5, 7, 2, 8, 1, 3, 4, 6\}$. In the "moving block" permutation $\pi : \{1, 2, 4, 5, 6, 7, 8\} \mapsto \{8, 1, 2, 3, 4, 5, 6, 7\}$. In the "i.i.d. block" permutation $\pi : \{\{1, 2\}, \{3, 4\}, \{5, 6\}, \{7, 8\}\} \mapsto \{\{3, 4\}, \{7, 8\}, \{1, 2\}, \{5, 6\}\}$, swapping all 2-blocks. The collection of all permutations forms the "i.i.d." group Π_{all} and the collection of all moving block permutations forms the "moving" group $\Pi_{\rightarrow}$, the collection of all "i.i.d." block permutations forms the "i.i.d. block" group Π_{mb}. The concept "group" formally includes the requirement that $\Pi\pi = \Pi$ for all $\pi \in \Pi$.

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

Definition of p-Value. The estimated p-value is

$$\hat{p} = 1 - \hat{F}\left(S(\hat{u})\right), \tag{2}$$

where

$$\hat{F}(x) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \mathbf{1}\{S(\hat{u}_\pi) < x\}.$$

Remark 3. We note that if the estimator used in approximating P_t^N is invariant to permutations of the data $\{Z_t\}$ across the time series dimension (which is the case for several of the estimators we consider in Sections 2.3 and 2.4), permuting $\{\hat{u}_t\}$ is equivalent to permuting $\{Z_t\}$.

We are often interested in testing pointwise hypotheses about α_t for $t \in \{T_0+1, \ldots, T\}$,

$$H_0 : \alpha_t = \alpha_t^o, \tag{3}$$

and in constructing pointwise confidence intervals for α_t. Hypothesis (3) can be tested by defining the data under the null as $\mathbf{Z} = (Z_1, \ldots, Z_{T_0}, Z_t)'$, provided that P_t^N can be estimated based on $\mathbf{Z}$. For $t \in \{T_0+1, \ldots, T\}$, pointwise $(1-\alpha)$ confidence intervals for α_t, $\mathcal{C}_{1-\alpha}(t)$, can be constructed via test inversion as described in Algorithm 1.

Algorithm 1 (Confidence Intervals for single periods). *(i) Choose a fine grid of G candidate values $\mathcal{A}_t = \{a_{1t}^o, \ldots, a_{Gt}^o\}$. (ii) For $a_t^o \in \mathcal{A}_t$, define $\mathbf{Z}$ for the null hypothesis $H_0 : \alpha_t = a_t^o$ and compute the corresponding p-value, $\hat{p}(a_t^o)$, using (2). (iii) Return the $(1-\alpha)$ confidence set $\mathcal{C}_{1-\alpha}(t) = \{a_t^o \in \mathcal{A}_t : \hat{p}(a_t^o) > \alpha\}$.*

2.3 Models for Counterfactual Proxies via Synthetic Control and Panel Data

The availability of control units motivates several modeling strategies for P_t^N (a non-exhaustive list of references on these different approaches is provided in the introduction).

2.3.1 Difference-in-Differences Methods

The difference-in-differences model postulates the following model for the counterfactual mean proxy (e.g., Doudchenko and Imbens, 2016, Section 5.1):[5]

$$P_t^N = \mu + \frac{1}{J} \sum_{j=2}^{J+1} Y_{jt}.$$

[5] This formulation assumes that all J control units are used to identify and estimate the counterfactual. In practice, researchers typically select the J controls from a larger donor pool of potential controls (this amounts to setting the weights of the controls which are not selected to zero). Note that if only one control unit is chosen, the counterfactual mean proxy is given by $P_t^N = \mu + Y_{2t}$.

This model automatically embeds the identifying information. The counterfactual mean proxy can be estimated as

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

2.3.2 Synthetic Control and Constrained Least Squares Estimators

The canonical synthetic control (SC) method (e.g., Abadie et al., 2010, 2015) postulates the following model:

$$P_t^N = \sum_{j=2}^{J+1} w_j Y_{jt}, \text{ where } w \geq 0 \text{ and } \sum_{j=2}^{J+1} w_j = 1. \tag{4}$$

We need to impose an identification condition that allows us to identify the weights w, for example:[6]

(SC) Assume that the structural shocks u_t for the treated units are uncorrelated with contemporaneous values of the outcomes, namely:

$$E\left(u_t Y_{jt}\right) = 0 \qquad \text{for} \qquad 2 \leq j \leq J+1, \tag{5}$$

The counterfactual is estimated as

$$\hat{P}_t^N = \sum_{j=2}^{J+1} \hat{w}_j Y_{jt}$$

We focus on the following canonical SC estimator for w:[7]

$$\hat{w} = \arg\min_{w} \sum_{t=1}^{T}\left(Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt}\right)^2 \text{ s.t. } w \geq 0 \text{ and } \sum_{j=2}^{J+1} w_j = 1. \tag{6}$$

[6] More generally, other exclusion restrictions and identifying assumptions could be used.

[7] This formulation of canonical SC without covariates is due to Doudchenko and Imbens (2016), who refer to the estimator (6) as constrained regression. We focus on the canonical problem (6) for concreteness. Abadie et al. (2010, 2015) consider a more generalized version, which also includes covariates into the estimation of the weights.

As an alternative, we can consider the more flexible model[8]

$$P_t^N = \mu + \sum_{j=2}^{J+1} w_j Y_{jt}, \text{ where } \|w\|_1 \leq 1, \tag{7}$$

maintaining the same identifying assumption (SC). The counterfactual is estimated as

$$\hat{P}_t^N = \hat{\mu} + \sum_{j=2}^{J+1} \hat{w}_j Y_{jt}$$

by the ℓ_1-constrained least squares estimator, or constrained Lasso, (e.g., Raskutti et al., 2011):

$$(\hat{\mu}, \hat{w}) = \arg\min_{(\mu, w)} \sum_{t=1}^{T} \left(Y_{1t} - \mu - \sum_{j=2}^{J+1} w_j Y_{jt} \right)^2 \text{ s.t. } \|w\|_1 \leq 1. \tag{8}$$

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 very weak assumptions.

Constrained Lasso encompasses both difference-in-differences and canonical SC as special cases (difference-in-differences is nested by setting $w = (1/J, \dots, 1/J)$, SC is nested by setting $\mu = 0$ and $w \geq 0$) and thus provides a unifying approach to the regression-based estimation of counterfactuals.

We will provide 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 (6) and (8).

[8] The 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 any formal properties of these estimators. Here we emphasize another version of relaxing SC, model (7), which leads to constrained Lasso (8). Constrained Lasso has really excellent theoretical and practical performance: it is tuning-free, performs very well empirically and in simulations, 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 the validity of our inference method under misspecification. It should be emphasized that this estimator generally differs from the cross-validated Lasso estimator.

2.3.3 Penalized Regression Methods

Consider the following linear model for P_t^N:

$$P_t^N = \mu + \sum_{j=2}^{J+1} w_j Y_{jt}.$$

Here, we maintain the identifying assumption (SC). Under this assumption the counterfactual is estimated by

$$\hat{P}_t^N = \hat{\mu} + \sum_{j=2}^{J+1} \hat{w}_j Y_{jt},$$

where

$$(\hat{\mu}, \hat{w}) = \arg\min_{(\mu,w)} \sum_{t=1}^{T} \left(Y_{1t} - \mu - \sum_{j=2}^{J+1} w_j Y_{jt} \right)^2 + \mathcal{P}(w), \tag{9}$$

and $\mathcal{P}(w)$ is a penalty function, which 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 (9).

Different variants of $\mathcal{P}(w)$ can be considered. For example:

- Lasso (Tibshirani, 1996): $\mathcal{P}(w) = \lambda \|w\|_1$, where λ is a tuning parameter. A version is the Post-Lasso estimator, which refits the weights after removing variables with zero weight.

- Elastic Net (Zou and Hastie, 2005): $\mathcal{P}(w) = \lambda\left((1-\alpha)\|w\|_2^2 + \alpha\|w\|_1\right)$, where λ and α are tuning parameters.

- Lava (Chernozhukov et al., 2017): $\mathcal{P}(w) = \inf_{a+b=w} \lambda\left((1-\alpha)\|a\|_2^2 + \alpha\|b\|_1\right)$, where 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. (2017), 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 fixed effects (FE) model for treated and untreated units:

$$Y_{jt}^{N} = \lambda_j' F_t + X_{jt}'\beta + u_{jt}, \quad \text{for } 1 \leq j \leq J+1 \text{ and } 1 \leq t \leq T, \tag{10}$$

where F_t are unobserved factors, λ_j are unit-specific factor loadings, and β is a vector of common coefficients.

(FE) We 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:

$$P_t^{N} = \lambda_1' F_t + X_{1t}'\beta. \tag{11}$$

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 (10); see e.g. Bai (2009) and Hansen and Liao (2016) for a version with high-dimensional covariates.

Model (10) nests the classical factor model

$$\lambda_j' F_t + \underbrace{X_{jt}'\beta}_{=0} = \lambda_j' F_t$$

and also covers the traditional linear FE model, in which

$$\lambda_j' F_t = \lambda_j + F_t.$$

There is a large body of work on these type of models. In econometrics these models are called interactive effects and augmented factor models; in statistics and machine learning they are called low-rank approximations and typically estimated through nuclear norm penalization methods or through universal singular value thresholding (upon imputing the missing entries with some reasonable proxies). A common application of these methods is the basic recommender system (e.g., the Netflix problem).

Hsiao et al. (2012) appears to the the first work that proposed the use of factor models for predicting the (missing) counterfactual response 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 fixed effects and factor models in SC designs are developed in Chan and Kwok (2016) and Li (2018).[9]

Other recent applications to predicting counterfactual responses include Amjad et al. (2017) and Athey et al. (2017) (using, respectively, the universal singular value thresholding and the nuclear norm penalization), albeit no inference methods are provided.[10] 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.[11] We shall be focusing on Bai (2009)'s alternating least squares estimator[12] and on matrix completion via nuclear norm penalization when verifying our conditions.

[9] Factor models are widely used in macroeconomics for causal inference and prediction; see, e.g., Stock and Watson (2016) and the references therein. In microeconometrics factor models are used for estimation of treatment/structural effects; see, e.g., Hansen and Liao (2016) where interactive fixed effects model are used to estimate the treatment effects of gun ownership on violence and the effect of institutions on growth.

[10] Note that Athey et al. (2017)'s analysis applies to a broader collection of problems with data missing in triangular patterns, nesting SC and difference-in-difference problems as special cases.

[11] Note 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 fixed effects model to estimate the model for the entire data under the null hypothesis.

[12] We 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.

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 treatment. For example, consider the following autoregressive model:[13]

$$\left.\begin{array}{l} Y_{1t}^N - \mu = \rho(Y_{1(t-1)}^N - \mu) + u_t \\ Y_{1t}^I - \mu = \rho(Y_{1(t-1)}^N - \mu) + \alpha_t + u_t \end{array}\right| \quad E(u_t) = 0, \quad \{u_t\} \text{ i.i.d.}, \quad t = 1, \ldots, T. \tag{12}$$

In this model 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_{1t}^I beyond the current values.[14] Under the null hypothesis, we can impute the unobserved counterfactual as $Y_{1t}^N = Y_{1t} - \alpha_t$, for $t > T_0$, 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:

$$\rho(\cdot) = \sum_{k=0}^{K} \rho_k \mathrm{L}^k(\cdot),$$

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,

$$\rho(\cdot) = m(\cdot, \mathrm{L}^1(\cdot), \ldots, \mathrm{L}^{\mathrm{k}}(\cdot)),$$

which arises in the context of using neural networks for predictive time series modeling (e.g., Chen and White, 1999; Chen et al., 2001) and we refer to the latter for identifying conditions.

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

[14] We leave the model with persistent, feed-forward effects, of the type $Y_{1t}^I = \rho(Y_{1(t-1)}^I) + \alpha_t + u_t$, to future work.

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:

$$\begin{array}{l|l|} Y_{1t}^N = C_t^N + \varepsilon_t & \varepsilon_t = \rho(\varepsilon_{t-1}) + u_t, \{u_t\} \text{ i.i.d. } E(u_t) = 0, \\ Y_{1t}^I = C_t^N + \alpha_t + \varepsilon_t & \{u_t\} \text{ is independent of } \{C_t^N\}, \end{array} \quad t = 1, \dots, T, \tag{13}$$

where C_t^N is a panel model proxy for Y_{1t}^N, identified by one of the panel data methods. Note that the model has the autoregressive formulation:

$$Y_{1t}^N = C_t^N + \rho(Y_{1(t-1)}^N - C_{t-1}^N) + u_t,$$

thereby generalizing the previous model.

Here the mean unbiased proxy for Y_{1t}^N is given by

$$P_t^N = C_t^N + \rho(\varepsilon_{t-1}).$$

P_t^N is a better proxy than C_t^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_t^N for counterfactuals with a time series model for the stochastic shock model in a nice way: we can identify C_t^N under the null by ignoring the time series structure, and then identify the time-series structure of the residuals $Y_{1t}^N - C_t^N$, where the missing observations Y_{1t}^N for $t > T_0$ are obtained as $Y_{1t}^N = Y_{1t} - \alpha_t$. Estimation can proceed analogously. This can improve upon the quality of our inferential procedure.

3 Theory

In this section, we provide theoretical justification for our conformal inference method. Our theoretical results are non-asymptotic in nature and hence hold in *finite samples*. When strong assumptions are imposed, the proposed approach is exact in a *model-free* manner. Under very weak assumptions, finite-sample bounds are provided for the size properties of our procedure; these bounds imply that our approach is asymptotically exact. In essence, we show that the proposed method automatically exploits exchangeability of the data (Section 3.1), accuracy of the estimator (Section 3.2.1) and stability of the estimator (Section 3.2.2).

3.1 Model-Free Exact Validity under Strong Assumptions on the Data

The following result shows that our conformal inference approach achieves exact finite sample size control if the estimated residuals, $\{\hat{u}_t\}$, are exchangeable. Importantly, this result is model-free in the sense that we do not need to use a correct or consistent estimator $\hat{P}_t^N$ for P_t^N. As a consequence, our procedure controls size under arbitrary forms of misspecification and is fully robust against overfitting.

Theorem 1 (Exact Validity). *Suppose that the Counterfactual Model stated in Assumption 1 holds. Let Π be $\Pi_{\rightarrow}$, Π_{all} or Π_{mb}. More generally, let Π form a group in the sense that $\Pi\pi = \Pi$ for all $\pi \in \Pi$. Suppose that $\{\hat{u}_t\}_{t=1}^T$ is exchangeable with respect to Π under the null hypothesis. Then, under the null hypothesis, the permutation p-value is unbiased in level:*

$$P\left(\hat{p} \leq \alpha\right) \leq \alpha.$$

Moreover, if $\{S(\hat{u}_\pi)\}_{\pi \in \Pi}$ has a continuous joint distribution,

$$\alpha - \frac{1}{|\Pi|} \leq P\left(\hat{p} \leq \alpha\right).$$

Exchangeability of the residuals is implied, for example, if the data $\{Z_t\}_{t=1}^T$ are i.i.d. or exchangeable under the null, as shown in Lemma 1, but holds more generally. For example, in the difference-in-difference model the outcome data can have an arbitrary common trend eliminated by differencing, making it possible for $\hat{u}_t = \hat{P}_t^N - P_t^N$ to be i.i.d. or exchangeable with non i.i.d. data.

Lemma 1 (Exchangeability with i.i.d. Data)**.** *Suppose that $\hat{u}_t = g(Z_t, \hat{\beta})$, where the estimator $\hat{\beta} = \hat{\beta}(\{Z_t\}_{t=1}^T)$ is invariant with respect to any permutation of the data. Then, if $\{Z_t\}_{t=1}^T$ is an i.i.d. or an exchangeable sequence, $\{\hat{u}_t\}_{t=1}^T$ is an exchangeable sequence.*

Exchangeability is a strong assumption and there are empirical settings where it is not plausible. In Section 3.2, we show that, under weak conditions, our inference procedure is approximately valid in general times series settings where exchangeability fails.

Finally, we would like to emphasize that the finite-sample validity under exchangeability highlights the robustness of our proposal. Since many machine learning methods and high-dimensional models might have overfitting in small samples, our finite-sample validity result implies that the proposed method does not suffer from overfitting or model misspecification. The fundamental reason is that we exploit symmetry of the procedure rather than completely relying on the accuracy of the estimator. For example, in Lemma 1, we essentially require that the null hypothesis be imposed when estimating the model.

This leads to invariance of the estimator under permutation and thus exchangeable residuals, which is why the proposed procedure is more robust than estimating the model only on the pretreatment data.

3.2 Approximate Validity under Weak Assumptions on the Data and the Estimators

In this section, we show that the proposed inference procedure achieves uniform approximate size control when the residuals are not exchangeable. We consider two different sets of conditions. In Section 3.2.1, we establish the approximate validity of our procedure for settings where 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 3.2.2, we show that if the data are stationary and weakly dependent, our procedure is approximately valid, provided that the estimator satisfies a certain stability condition. Importantly, this stability condition does not require correct specification of P_t^N nor consistency of the estimator of P_t^N.

3.2.1 Approximate Validity under Estimator Consistency

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

Assumption 2 (Regularity of the Stochastic Shock Process)**.** *Assume that the p.d.f 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 i.i.d., or*

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

We can view Assumption 2 as much weaker than the previous assumptions, since the data can be non-stationary and exhibit general dependence patterns. Assumption 2.1 of i.i.d. shocks is our first sufficient condition. Under this condition, we will be able to use i.i.d. permutations, giving us a precise estimate of the p-value. The i.i.d. assumption can be replaced by Assumption 2.2, which is a widely accepted, weak condition, holding for many commonly encountered stochastic processes. 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 4. The assumption above can be generalized further, by requiring that the stochastic process $\{u_t\}_{t=1}^T$ satisfies one of the following conditions conditional on a random element V:

1. *Exchangeability: $\{u_t\}$ are i.i.d. variables, conditional on V, or*
2. *Conditional ergodicity: $\{u_t\}$ are stationary, strongly mixing, conditional on V, with the sum of the mixing coefficient bounded by M.*

Remark 5. 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 certain "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. We leave the development of concrete proposals for modeling heteroscedasticity to future research.

We also impose the following condition on the estimation error under the null hypothesis.

Assumption 3 (Consistency of the Counterfactual Estimators under the Null). *Let there be sequence 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 5, we provide explicit primitive conditions as well as references to explicit primitive conditions, which imply Assumption 3.

Theorem 2 (Approximate Validity under Consistent Estimation). *We assume that T_* is fixed and that $T \to \infty$. Suppose that the Counterfactual Model stated in Assumption 1 holds, and that Assumption 3 holds. Impose Assumption 2.1 if i.i.d. permutations are used. Impose Assumption 2.2, if moving block permutations are used. Assume the statistic $S(u)$ has a density function bounded by D under the null. Then, under the null hypothesis, the p-value is approximately unbiased in size:*

$$|P(\hat{p} \leq \alpha) - \alpha| \leq C(\tilde{\delta}_T + \delta_T + \sqrt{\delta_T} + \gamma_T) \to 0,$$

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

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 have, subject to assumed invariance to the policy. The smaller the variance of the shocks, the more power the testing procedure will have.

3.2.2 Approximate Validity under Estimator Stability

In practice, consistency of the estimators for the counterfactual mean proxies P_t^N may be questionable in certain settings. In this section, we therefore consider a notion of approximate exchangeability, which only requires the estimator to be *stable* instead of consistent. Specifically, we show that our conformal inference approach is approximately valid if the estimator $\hat{P}_t^N$ satisfies a certain perturbation stability condition, which we formally introduce below. We would like to emphasize that this stability condition does not require the estimator to be consistent nor does it rely on correct specification of the counterfactual mean proxies.

The basic idea 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 permutation approximately has a uniform distribution under the null. This is a simple consequence of 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 $S(\mathbf{Z},\beta)=\phi(Z_{T_0+1},,\ldots,Z_{T_0+T_*};\beta)$. Let $\{\tilde{Z}_t\}_{t=1}^T$ be i.i.d. from the distribution of Z_1 and independent of $\mathbf{Z}$. For any $H\subset\{1,\ldots,T\}$, 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 $\{\tilde{Z}_t\}_{t\in H}$. We impose stability under a class of H.

Let $R\in\mathbb{N}$ and define $m=\lfloor T_0/R\rfloor$. For $j\in\{1,\ldots,R\}$, let $H_j=\{(j-1)m+1,\ldots,jm\}$. Let $k\in\mathbb{N}$ satisfy $T_*<k<m$. We let $\widetilde{H}_j$ denote the k-enlargement of H_j, i.e., $\widetilde{H}_j=\{s:\min_{t\in H_j}|s-t|\leq k\}$. It is not hard to see that $\widetilde{H}_j=\{(j-1)m+1-k,\ldots,jm+k\}$ for $2\leq j\leq R-1$, $\widetilde{H}_1=\{1,\ldots,m+k\}$ and $\widetilde{H}_R=\{(R-1)m+1-k,\min\{Rm+k,T\}\}$.

Let $\Psi(x;\beta)=P(\phi(Z_{T_0+1},\ldots,Z_{T_0+T_*};\beta)\leq x)$. Our strategy is to show that, under the null hypothesis, $\hat{F}(\phi(Z_{T_0+1},\ldots,Z_{T_0+T_*};\hat{\beta}(\mathbf{Z})))$ approximately has a uniform distribution on $(0,1)$. The theoretical argument exploits the stability condition below and shows that $\hat{F}(\phi(Z_{T_0+1},\ldots,Z_{T_0+T_*};\hat{\beta}(\mathbf{Z})))$ can be approximated by $\Psi\left(\phi(\bar{Z}_{T_0+1},\ldots,\bar{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R}));\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)$, where $(\bar{Z}_{T_0+1},\ldots,\bar{Z}_{T_0+T_*})$ has the same distribution as $(Z_{T_0+1},\ldots,Z_{T_0+T_*})$ and is independent of $\mathbf{Z}_{\widetilde{H}_R}$. This essentially confirms the above intuition that for stable estimators, $\hat{\beta}(\mathbf{Z})$ is almost independent of the last few observations $(Z_{T_0+1},\ldots,Z_{T_0+T_*})$.

Assumption 4 (Estimator Stability)**.** *Let $\Pi=\Pi_{\rightarrow}$. There exist increasing 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

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

for any $H\in\{\widetilde{H}_1,\ldots,\widetilde{H}_R\}$, where $\dot{\mathbf{Z}}\overset{d}{=}\mathbf{Z}$ and $\dot{\mathbf{Z}}$ is independent of $(\mathbf{Z},\{\tilde{Z}_t\}_{t=1}^T)$.

Assumption 4 specifies the estimator stability condition. This is similar to the perturb-one sensitivity of Lei et al. (2018); see their 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. Sufficient conditions for Assumption 4 are provided in Section 6. Next, we impose additional regularity conditions on the data.

Assumption 5 (Regularity of the Data)**.** *$\{Z_t\}_{t=1}^T$ is 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 $D_1,D_2,D_3>0$. There exist sequences $\xi_T>0$ and $\gamma_{2,T}=o(1)$ such that $P\left(\sup_{x\in\mathbb{R}}\left|\partial\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)/\partial x\right|\leq\xi_T\right)\geq 1-\gamma_{2,T}$ for $1\leq j\leq R$.*

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, $D_3=1$; see Carrasco and Chen (2002). Let $(\dot{Z}_{T_0+1},\ldots,\dot{Z}_{T_0+T_*})$ be an independent copy of $(Z_{T_0+1},\ldots,Z_{T_0+T_*})$ and also independent of $(\mathbf{Z},\{\tilde{Z}\}_{t=1}^T)$. The bounded derivative of $\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)$ condition says that the p.d.f of $\phi(\dot{Z}_t,\ldots,\dot{Z}_{t+T_*-1};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j}))$ conditional on $\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})$ 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, the p.d.f of $Y_t-X_t'w$ is bounded by a constant that does not depend on w.

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

Theorem 3 (Approximate Validity under Estimator Stability). *Suppose that the Counterfactual Model stated in Assumption 1 holds, and 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 $R < T_0/2$,*

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

In the theoretical arguments, we actually show a stronger result: the above bound holds for $E|P(\hat{p} \leq \alpha \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})) - \alpha|$. Since the stability condition states that $\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R}) \approx \hat{\beta}(\mathbf{Z})$, this means that $\hat{p}$ conditional on $\hat{\beta}(\mathbf{Z})$ almost has a uniform distribution on $(0, 1)$; in the case of i.i.d or exchangeable data, $\hat{p}$ conditional on $\hat{\beta}(\mathbf{Z})$ has an exact uniform distribution. For this reason, we can view Theorem 3 as a result for approximate exchangeability.

Due to the exponential decay of $\beta_{\text{mixing}}(\cdot)$, the bound in Theorem 3 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 $|\widetilde{H}_j| = \lfloor T_0/R \rfloor + 2k$, Assumption 5 only requires that the changes to $S(\mathbf{Z}^{\pi}, \hat{\beta}(\mathbf{Z}))$ is 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. Notice that R is only needed in the theoretical arguments; we do not need to choose R when implementing the proposed procedure.

4 Extensions

In this section, we develop three extensions of our main results.

4.1 Testing Hypotheses about Average Effects over Time

In addition to testing sharp null hypotheses, researchers are often also interested in testing hypotheses about average effects over time:

$$H_0 : \bar{\alpha} = \bar{\alpha}^o, \tag{14}$$

where

$$\bar{\alpha} = \frac{1}{T_*}\sum_{t=T_0+1}^{T} \alpha_t.$$

To simplify the exposition, we assume that T/T_* is an integer. Our inference procedure can modified to test hypothesis (14). Towards this end, note that Assumption 1 implies the following model for the average potential outcomes $\bar{Y}_{1r}^N = T_*^{-1}\sum_{t=r}^{r+T_*-1} Y_t^N$ and $\bar{Y}_{1r}^I = T_*^{-1}\sum_{t=r}^{r+T_*-1} Y_t^I$:

$$\left.\begin{array}{l} \bar{Y}_{1r}^N = \bar{P}_r^N + \bar{u}_r \\ \bar{Y}_{1r}^I = \bar{P}_r^N + \bar{\alpha}_r + \bar{u}_r \end{array}\right| \quad E(\bar{u}_r) = 0, \quad r = 1, T_* + 1, \dots, T_0 + 1,$$

where $\bar{P}_r^N = T_*^{-1}\sum_{t=r}^{r+T_*-1} P_t^N$ and $\bar{u}_r = T_*^{-1}\sum_{t=r}^{r+T_*-1} u_t$. Define the aggregated data under the null as $\bar{\mathbf{Z}} = (\bar{Z}_1, \dots, \bar{Z}_{T_0+1})'$, where

$$\bar{Z}_r = \begin{cases} \left(\bar{Y}_{1r}^N, \bar{Y}_{2r}^N, \dots, \bar{Y}_{J+1r}^N, \bar{X}_{1r}', \dots, \bar{X}_{J+1r}'\right)', & r < T_0 + 1 \\ \left(\bar{Y}_{1r}^N - \bar{\alpha}^o, \bar{Y}_{2r}^N, \dots, \bar{Y}_{J+1r}^N, \bar{X}_{1r}', \dots, \bar{X}_{J+1r}'\right)', & r = T_0 + 1 \end{cases}$$

and $\bar{X}_{jr} = T_*^{-1}\sum_{t=r}^{r+T_*-1} X_{jt}$ for $j = 1, \dots, J+1$. Note that testing hypothesis (14) is equivalent to testing the simple hypothesis (3) based on the aggregated data $\bar{\mathbf{Z}}$. Specifically, we compute the estimated average proxy $\hat{\bar{P}}_r^N$ based on the aggregated data $\bar{\mathbf{Z}}$ and obtain the residuals

$$\hat{\bar{u}} = (\hat{\bar{u}}_1, \hat{\bar{u}}_{T_*+1}, \dots, \hat{\bar{u}}_{T_0+1}), \quad \hat{\bar{u}}_r = \bar{Y}_{1r}^N - \hat{\bar{P}}_r^N, \quad r = 1, T_* + 1, \dots, T_0 + 1.$$

The test statistic is $S\left(\hat{\bar{u}}\right)$ and p-values can be obtained based on permutations of $(\hat{\bar{u}}_1, \hat{\bar{u}}_{T_*+1}, \dots, \hat{\bar{u}}_{T_0+1})$ as described in Section 2.2. The formal properties of the test follow from the results in Section 3. The key assumption underlying this procedure is that the average mean proxy $\bar{P}_r^N$ can be identified and estimated based on the aggregated data $\bar{\mathbf{Z}}$. This assumption is often satisfied if the model for P_t^N is linear. Furthermore, note that the effective sample size is T/T_* instead of T and, consequently, T needs to be substantially larger than T_*.

4.2 Multiple Treated Units

In the main part of this paper, we focus on an aggregate panel data setting with only one treated unit. Here we briefly discuss how our method can be extended to accommodate multiple treated units. Consider a setup with L treated units, indexed by $j = 1, \dots, L$, and J control units, indexed by $j = L + 1, \dots, J + N$. Suppose that Assumption 1 holds for all treated units:

$$\left.\begin{array}{l} Y_{jt}^N = P_{jt}^N + u_{jt} \\ Y_{jt}^I = P_{jt}^N + \alpha_{jt} + u_{jt} \end{array}\right| \quad E(u_{jt}) = 0, \quad t = 1, \dots, T, \quad j = 1, \dots, L.$$

Under this assumption, hypotheses about the unit-specific treatment effects $\{\alpha_{jt}\}$ can be tested by applying the proposed inference procedure unit-by-unit. In addition, one is often also interested in conducting inference about the average treatment effects across units $\{\bar{\alpha}_t\}$, where

$$\bar{\alpha}_t = \frac{1}{L}\sum_{j=1}^{L} \alpha_{jt}.$$

Specifically, consider the following null hypothesis:

$$H_0 : (\bar{\alpha}_{T_0+1}, \dots, \bar{\alpha}_T) = \left(\bar{\alpha}^o_{T_0+1}, \dots, \bar{\alpha}^o_T\right). \tag{15}$$

To test hypothesis (15), note that if Assumption 1 holds for all treated units, we have the following model for the average potential outcomes $\bar{Y}_t^N = L^{-1}\sum_{j=1}^{L} Y_{jt}^N$ and $\bar{Y}_t^I = L^{-1}\sum_{j=1}^{L} Y_{jt}^I$:

$$\left.\begin{array}{l} \bar{Y}_t^N = \bar{P}_t^N + \bar{u}_t \\ \bar{Y}_t^I = \bar{P}_t^N + \bar{\alpha}_t + \bar{u}_t \end{array}\right| \quad E(\bar{u}_t) = 0, \quad t = 1, \dots, T,$$

where $\bar{P}_t^N = L^{-1}\sum_{j=1}^{L} P_{jt}^N$ and $\bar{u}_t = L^{-1}\sum_{j=1}^{L} u_{jt}$. Define the data under the null as $\bar{\mathbf{Z}} = (\bar{Z}_1, \dots, \bar{Z}_T)'$, where

$$\bar{Z}_t = \begin{cases} \left(\bar{Y}_t^N, Y_{L+1t}^N, \dots, Y_{J+Nt}^N, \bar{X}_t', X_{L+1t}', \dots, X_{J+Lt}'\right)', & t \leq T_0 \\ \left(\bar{Y}_{1t} - \bar{\alpha}_t^o, Y_{L+1t}^N, \dots, Y_{J+Lt}^N, \bar{X}_t', X_{L+1t}', \dots, X_{J+Lt}'\right)', & t > T_0, \end{cases}$$

and $\bar{X}_t = L^{-1}\sum_{j=1}^{L} X_{jt}$. To test hypothesis (15), we compute the estimated average proxy $\hat{\bar{P}}_t^N$ based on the aggregated data $\bar{\mathbf{Z}}$ and obtain the residuals

$$\hat{\bar{u}} = (\hat{\bar{u}}_1, \dots, \hat{\bar{u}}_T), \quad \hat{\bar{u}}_t = \bar{Y}_t^N - \hat{\bar{P}}_t^N, \quad t = 1, \dots, T.$$

The test statistic is $S\left(\hat{\bar{u}}\right)$ and p-values can be obtained based on permutations of $(\hat{\bar{u}}_1, \dots, \hat{\bar{u}}_T)$ as described in Section 2.2. The formal properties of this test follow from the results in Section 3.

4.3 Placebo Specification Tests

Here we consider an easy-to-implement placebo specification test for assessing the validity of the key assumptions underlying our approach. The idea is to test the null hypothesis

$$H_0 : \alpha_{T_0-\tau+1} = \cdots = \alpha_{T_0} = 0, \tag{16}$$

for a given $\tau \geq 1$ based on the pre-treatment data $\tilde{\mathbf{Z}} = (Z_1, \ldots, Z_{T_0})'$. Using an appropriate CSC method, we compute the counterfactual mean proxies $\hat{P}_t^N$ based on $\tilde{\mathbf{Z}}$ and obtain the residuals

$$\hat{u} = (\hat{u}_1, \ldots, \hat{u}_{T_0})', \quad \hat{u}_t = Y_{1t}^N - \hat{P}_t^N, \quad t = 1, \ldots, T_0.$$

We then apply the proposed inference method, treating $\{1, \ldots, T_0 - \tau\}$ as the pre-treatment period and $\{T_0 - \tau + 1, \ldots, T_0\}$ as the post-treatment period. If the assumptions underlying our inference procedure are correct, the null hypothesis (16) is true. A rejection of hypothesis (16) thus provides evidence against correct specification. The theoretical properties of such specification tests follow directly from the results in Section 3.

5 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 approximate validity of the proposed procedure. In particular, these conditions can be used to verify Assumption 3.

5.1 Difference-in-Differences

In Section 2.3.1, we have seen that under the canonical difference-in-differences models, the counterfactual proxy is given as

$$P_t^N = \mu + \frac{1}{J} \sum_{j=2}^{J+1} Y_{jt},$$

We consider the following estimator for the counterfactual:

$$\hat{P}_t^N = \hat{\mu} + \frac{1}{J} \sum_{j=2}^{J+1} Y_{jt},$$

where

$$\hat{\mu} = \frac{1}{T} \sum_{t=1}^{T} \left(Y_{1t}^N - \frac{1}{J} \sum_{j=2}^{J+1} Y_{jt} \right) = \mu + \frac{1}{T} \sum_{t=1}^{T} u_t.$$

Since $\hat{P}_t^N - P_t^N = \hat{\mu} - \mu$, 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.

5.2 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 denote a generic vector of outcomes and X denote the design matrix throughout this section. For example, in Section 2, we set $Y = Y_1^N$ and $X = (Y_2^N, \ldots, Y_{J+1}^N)$, where $Y_j^N = (Y_{j1}^N, \ldots, Y_{jT}^N)' \in \mathbb{R}^T$ for $1 \leq j \leq J+1$. These models can be written as

$$Y = Xw + u, \tag{17}$$

where $u = (u_1, u_2, \ldots, u_T)' \in \mathbb{R}^T$. Identification is achieved by requiring that X and u be uncorrelated (cf. condition (SC)).

Under the framework in (17), 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

$$\hat{w} = \arg\min_{w} \ \|Y - Xw\|_2 \quad \text{s.t. } w \in \mathcal{W} = \{v \geq 0, \|v\|_1 = 1\}. \tag{18}$$

The constrained Lasso estimator is

$$\hat{w} = \arg\min_{w} \|Y - Xw\|_2 \quad \text{s.t. } w \in \mathcal{W} = \{v : \|v\|_1 \leq K\}, \tag{19}$$

where $K > 0$ is a tuning parameter. In light of the estimator (18), 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.[15]

Lemma 2 (Constrained Least Squares Estimators). *Consider*

$$\hat{w} = \arg\min_{v} \ \|Y - Xv\|_2 \quad \text{s.t. } v \in \mathcal{W},$$

[15] To simplify the exposition, we do not include an intercept in Lemma 2. Similar arguments could be used to prove an analogous result with an unconstrained intercept.

where $\mathcal{W}$ is a subset of $\{v : \|v\|_1 \le K\}$ and K is bounded. Assume $w \in \mathcal{W}$, the data is $\beta-$mixing with exponential speed and other assumptions listed at the beginning of the proof, then the estimator enjoys the finite-sample performance bounds stated in the proof, in particular:

$$\frac{1}{T}\sum_{t=1}^{T}(\hat{P}_t^N - P_t^N)^2 = o_P(1) \quad \textit{and} \quad \hat{P}_t^N - P_t^N = o_P(1), \textit{ for any } T_0 + 1 \le t \le T.$$

Lemma 2 provides some 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 8. It is also important to note that the result in Lemma 2 does not require any sparsity assumptions on w, allowing for dense vectors. Moreover, compared to typical high-dimensional estimators (e.g., Lasso or Dantzig selector), our estimator does not require tuning parameters that can be difficult to choose in practice. Finally, we would like to emphasize that Lemma 2 provides new theoretical results for the canonical SC estimator in settings with potentially very many control units.

5.3 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 $N = J + 1$.

5.3.1 Pure Factor Models

Recall from Section 2.3.4 the standard large 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_{1t}^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 (Condition FE).

We use the standard principal component analysis (PCA) for estimating P_t^N.[16] Let

[16] Note that PCA amounts to singular value decomposition, which can be computed using polynomial

$Y^N \in \mathbb{R}^{T\times N}$ be the matrix whose (t,j) entry is Y_{jt}^N. We compute $\hat{F} = (\hat{F}_1, \ldots, \hat{F}_T)' \in \mathbb{R}^{T\times k}$ 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 3 (Factor/Matrix Completion Model)**.** *Assume standard regularity conditions given in Bai (2003) including the identification condition FE, listed at the beginning of the proof of this lemma. Consider the factor model and the principal component estimator. Then, for any $1 \le t \le T$, as $N \to \infty$ and $T \to \infty$*

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

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.

5.3.2 Factor plus Regression Model: Interactive Fixed Effects Model

Now we study the general form of panel models with interactive fixed effects. Following Section 2.3.4, these models take the form

$$Y_{jt}^N = \lambda_j' F_t + X_{jt}'\beta + u_{jt},$$

where $X_{jt} \in \mathbb{R}^{k_x}$ is observed covariates and $F = (F_1, \ldots, F_T)' \in \mathbb{R}^{T\times k}$ and $\Lambda = (\lambda_1, \ldots, \lambda_N)' \in \mathbb{R}^{N\times k}$ represent the k-dimensional unobserved factors and their loadings, respectively. The counterfactual proxy for Y_{1t}^N is $P_t^N = \lambda_1' F_t + X_{1t}'\beta$. 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.

The two most popular estimators are the common correlated effects (CCE) estimator by Pesaran (2006) and the iterative least squares estimator by Bai (2009). In this paper, we focus on the iterative least squares approach, but analogous results can be established for

time algorithms, (e.g., Trefethen and Bau III, 1997, Lecture 31).

CCE estimators. The notations for F_t, λ_j, $\hat{F}_t$ and $\hat{\lambda}_j$ are the same as before. We compute

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

The estimate for P_t^N is $\hat{P}_t^N = \hat{\lambda}_1'\hat{F}_t + X_{1t}'\hat{\beta}$. The following result states the validity of applying this estimator in conjunction with our inference method.

Lemma 4 (Interactive Fixed Effect Model)**.** *Assume the standard conditions in Bai (2009) including the identification condition FE. Then, for any $1 \leq t \leq T$,*

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

Note that under 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.

5.3.3 Matrix Completion via Nuclear Norm Regularization

Suppose that

$$Y_{jt}^N = M_{jt} + u_{jt}, \quad \text{for } 1 \leq j \leq J+1 \text{ and } 1 \leq t \leq T, \tag{20}$$

where M_{jt} is the (j,t)-element of an unknown matrix $M \in \mathbb{R}^{N \times T}$ satisfying $\|M\|_* \leq K$, where $\|\cdot\|_*$ denotes the nuclear norm, i.e., the sum of singluar values. We observe Y_{jt}^N for $(j,t) \in \{1,\ldots,T\} \times \{1,\ldots,J+1\} \backslash \{(1,t) : T_0 + 1 \leq t \leq T\}$. The identifying condition is that $E(u \mid M) = 0$ and that conditional on M, $\{u_j\}_{j=1}^N$ is independent across j, where $u_j = (u_{j1},\ldots,u_{jT})' \in \mathbb{R}^T$. 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 of matrix completion considers the model (20) under the assumption of missing at random and exploits the assumption that the rank of M is low; see for instance 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) among many others. Recently, Athey et al. (2017) 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 theoretical 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 $\{Y_{jt}^N : 1 \le j \le J+1,\ 1 \le t \le T\}$. The estimator for M we examine here is closely related to existing nuclear norm regularized estimators and is defined as

$$\hat{M} = \underset{A \in \mathbb{R}^{N \times T}}{\arg\min} \sum_{t=1}^{T} \sum_{j=1}^{N} (Y_{jt}^N - A_{jt})^2 \quad \text{s.t. } \|A\|_* \le K, \tag{21}$$

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 $\sqrt{NT}$. A clear guidance regarding how to choose K is still unavailable, but following Athey et al. (2017) one can use cross-validation.[17] 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 5. *Consider the estimator $\hat{M}$ defined in (21). Assume that $\|M\|_* \le K$. Let the conditions listed at the beginning of the proof hold. Then, for any $T_0 + 1 \le t \le T$,*

$$\hat{P}_t^N - P_t^N = o_P(1) \quad \textit{and} \quad \frac{1}{T} \sum_{t=K+1}^{T} \left(\hat{P}_t^N - P_t^N \right)^2 = o_P(1).$$

The result is notable because no sub-Gaussian assumptions are required. The estimator in (21) does not explicitly require a low-rank condition on M. Instead, we impose a growth restriction on K. In the case in which M is generated by a strong factor structure and the null hypothesis contain full information on the missing entries, we can choose $K \asymp \sqrt{NT}$ and our consistency result holds as long as $N, T \to \infty$ and $E(|u_{jt}|^{2+c} \mid M)$ is uniformly bounded for some $c > 0$. Notice that in the case of weak factors, we can choose $K \ll \sqrt{NT}$ and obtain consistency.

5.4 Time Series and Fused Models

As pointed out in Section 2.4, time series models, such as AR models, can be used to model counterfactual proxies with or without control units. We now discuss low-level conditions

[17] The properties of cross-validation remain unknown in these settings.

under which fitting these models yields estimates good enough for the purpose of our general conformal inference approach.

5.4.1 Autoregressive Models

The linear autoregressive model with K lags can be written as[18]

$$Y_{1t}^N = \rho_0 + \sum_{j=1}^{K} \rho_j Y_{1t-j}^N + u_t,$$

where $\{u_t\}_{t=1}^T$ is an i.i.d. sequence with $E(u_t) = 0$. Here the counterfactual proxy for Y_{1t}^N is $P_t^N = \rho_0 + \sum_{j=1}^K \rho_j Y_{1t-j}^N$. It will be useful to write P_t^N as $P_t^N = y_t' \rho$, where

$$y_t = (1, Y_{1t-1}^N, Y_{1t-2}^N, \dots, Y_{1t-K}^N)' \in \mathbb{R}^{K+1},$$

and $\rho = (\rho_0, \dots, \rho_K)' \in \mathbb{R}^{K+1}$. The coefficient vector ρ can be estimated using least squares:

$$\hat{\rho} = \left(\sum_{t=K+1}^{T} y_t y_t' \right)^{-1} \left(\sum_{t=K+1}^{T} y_t Y_{1t}^N \right).$$

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

Lemma 6 (Linear AR Model). *Suppose that $\{u_t\}_{t=1}^T$ is an i.i.d 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 $T_0 + 1 \leq t \leq T$,*

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

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. Thus, the counterfactual proxy is $P_t^N = \rho(Y_{1t-1}^N, Y_{1t-2}^N, \dots, Y_{1t-K}^N)$.

We allow ρ to be parametric, nonparametric or semi-parametric. In general, we only require a consistent estimator for ρ. Let $\hat{\rho}$ be an estimator for ρ and $\hat{P}_t^N = \hat{\rho}(Y_{1t-1}^N, Y_{1t-2}^N, \dots, Y_{1t-K}^N)$.

[18] Here the model seems different, but Section 2.4's model implies this one with $\rho_0 = \mu(1 - \sum_{j=1}^K \rho_j)$

Lemma 7 (Nonlinear AR Model). *Suppose that (1)* $\|\hat{\rho} - \rho\| = O_P(r_T)$ *with* $r_T = o(1)$ *for some appropriate norm* $\|\cdot\|$ *and* $\max_{K+1\le t\le T} |\hat{\rho}(Y_{1t-1}^N, Y_{1t-2}^N, \ldots, Y_{1t-K}^N) - \rho(Y_{1t-1}^N, Y_{1t-2}^N, \ldots, Y_{1t-K}^N)| \le \ell_T \|\hat{\rho} - \rho\|$ *for some* $\ell_T r_T = o(1)$. *Then, for any* $T_0 + 1 \le t \le T$,

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

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

5.4.2 Fused Panel/Time Series Models with AR Errors

Here, we provide generic conditions for 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 i.i.d errors. In Equation (13) of Section 2.4, we introduce an autoregressive structure in the error terms:

$$Y_{1t}^N = C_t^N + \varepsilon_t \qquad \text{and} \qquad \varepsilon_t = \rho(\varepsilon_{t-1}) + u_t,$$

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 $P_t^N = C_t^N + \rho(\varepsilon_{t-1})$.

The estimation for P_t^N is done 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 the estimated residuals $\{\hat{\varepsilon}_t\}_{t=1}^T$ to an autoregressive model, where $\hat{\varepsilon}_t = Y_{1t}^N - \hat{C}_t^N$. For simplicity, we consider a linear model in the second stage estimation but analogous results can be obtained for more general models. To be specific, assume that

$$\varepsilon_t = x_t'\rho + u_t,$$

where $x_t = (\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-K})' \in \mathbb{R}^K$ and $\rho = (\rho_1, \rho_2, \ldots, \rho_K)' \in \mathbb{R}^K$.

Given $\{\hat{\varepsilon}_t\}_{t=1}^T$ from the first-stage estimation, we define $\hat{x}_t = (\hat{\varepsilon}_{t-1}, \hat{\varepsilon}_{t-2}, \ldots, \hat{\varepsilon}_{t-K})' \in \mathbb{R}^K$ 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 8 (AR Errors). *Suppose that $\{u_t\}_{t=1}^T$ is an i.i.d 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) $\sum_{t=1}^T (\hat{C}_t^N - C_t^N)^2 = o_P(T)$, (2) $\hat{C}_t^N - C_t^N = o_P(1)$ for $T_0 - K + 1 \leq t \leq T$. Then, for any $T_0 + 1 \leq t \leq T$,*

$$\hat{P}_t^N - P_t^N = o_P(1) \quad \textit{and} \quad \sum_{t=K+1}^T \left(\hat{P}_t^N - P_t^N\right)^2 = o_P(T)$$

Note that the conditions in Lemma 8 for the autoregressive part are the same as in Lemma 6. The requirement on the consistency of $\hat{C}_t^N$ can be verified using existing results, e.g., those in Sections 5.1 – 5.3.

6 Sufficient Conditions for Estimator Stability

In this section, we provide sufficient conditions for the estimator stability Assumption 4. We first present a generic sufficient condition for low-dimensional models. For high-dimensional models, the theoretical analysis is much more difficult and a case-by-case analysis is needed. We are not aware of any theoretical work that establishes Assumption 4 for any high-dimensional model. Here we provide the first such result by verifying the stability condition for constrained Lasso.

6.1 Generic Sufficient Condition for Low-dimensional Models

Consider $\hat{\beta}(\mathbf{Z}) = \arg\min_{\beta \in \mathcal{B}} \hat{L}(\mathbf{Z}; \beta)$, where $\hat{L}(\mathbf{Z}; \beta)$ is a loss function and $\mathcal{B} \subset \mathbb{R}^p$ for a fixed p. Let $\mathcal{H}$ be a set of subsets of $\{1, \ldots, T\}$. Notice that Assumption 4 only requires $\mathcal{H}$ to be a singleton, but in this subsection and the next, we allow $\mathcal{H}$ to be a class of subsets.

Lemma 9. *Suppose that the following conditions hold:*
(1) $\sup_{\beta \in \mathcal{B}} |\hat{L}(\mathbf{Z}; \beta) - L(\beta)| = o_P(1)$ for some non-random $L(\cdot)$.
(2) $\max_{H \in \mathcal{H}} \sup_{\beta \in \mathcal{B}} |\hat{L}(\mathbf{Z}_H; \beta) - L(\beta)| = o_P(1)$.
(3) $L(\cdot)$ is continuous at β_*, $\min_\beta L(\beta)$ has a unique minimum at β_* and $\mathcal{B}$ is compact.
Then $\max_{H \in \mathcal{H}} \|\hat{\beta}(\mathbf{Z}) - \hat{\beta}(\mathbf{Z}_H)\|_2 = o_P(1)$.

In the literature of misspecified models, β_* is usually referred to as the pseudo-true value, e.g., White (1996). In M-estimation with $\hat{L}(\mathbf{Z}; \beta) = T^{-1} \sum_{t=1}^T l(Z_t; \beta)$, one can often show $\sup_\beta |\hat{L}(\mathbf{Z}; \beta) - L(\beta)| = o_P(1)$ with $L(\beta) = El(Z_1; \beta)$; in GMM models with $\hat{L}(\mathbf{Z}; \beta) = \|T^{-1} \sum_{t=1}^T \psi(Z_t; \beta)\|_2$, one can often use $L(\beta) = \|E\psi(Z_1; \beta)\|_2$.

The proof of Lemma 9 shows that $\|\hat{\beta}(\mathbf{Z}) - \beta_*\|_2 = o_P(1)$ and $\max_{H\in\mathcal{H}} \|\hat{\beta}(\mathbf{Z}_H) - \beta_*\|_2 = o_P(1)$. In other words, the stability of the estimator arises from the consistency to the pseudo-true value β_*. Such consistency holds under very weak conditions. We essentially only require a uniform law of large numbers. This can be verified for many low-dimensional models under weakly dependent data. For low-dimensional models, the conclusion of Lemma 9 translates to Assumption 4 once we derive a bound on $\sup_{\beta_1\neq\beta_2} |S(\mathbf{Z};\beta_1) - S(\mathbf{Z};\beta_2)|/\|\beta_1 - \beta_2\|_2$; this requires knowledge of the model structure.

6.2 Constrained Lasso

Here we propose sufficient conditions for estimator stability for constrained Lasso. In contrast to Sections 2.3.2 and 5.2, we do not impose correct specification but study the behavior of the constrained Lasso estimator under potential misspecification. To make this explicit, we use β instead of w to denote the coefficient vector in this subsection. Here, it is possible that $EX_t(Y_t - X_t'\beta) \neq 0$ for any $\beta \in \mathcal{W}$. In practice, this arises when the relationship between X_t and Y_t is not linear or when the constraint set $\mathcal{W}$ is too small. For example, the true parameter could be non-sparse with exploding ℓ_1-norm, e.g., $\beta = (1,\ldots,1)'/\sqrt{J}$.

We first introduce some additional notation. Define $Y_t = Y_{1t}^N$ and $X_t = (Y_{2t}^N, \ldots, Y_{J+1t}^N)$ and let $\{(\tilde{Y}_t, \tilde{X}_t)\}_{t=1}^T$ be i.i.d. from the distribution of (Y_1, X_1) and independent of the data $\{(Y_t, X_t)\}_{t=1}^T$. The constrained Lasso objective functions based on the data under the original data and after switching out observations with $t \in H$ are given by

$$\hat{Q}(\beta) = \frac{1}{T}\sum_{t=1}^T (Y_t - X_t'\beta)^2 \quad \text{and} \quad \hat{Q}_H(\beta) = T^{-1}\sum_{t=1}^T (Y_{t,H} - X_{t,H}'\beta)^2,$$

where $(Y_{t,H}, X_{t,H}) = (Y_t, X_t)$ for $t \notin H$ and $(Y_{t,H}, X_{t,H}) = (\tilde{Y}_t, \tilde{X}_t)$ for $t \in H$. The corresponding constrained Lasso estimators are

$$\hat{\beta}(\mathbf{Z}) = \arg\min_{\beta\in\mathcal{W}} \hat{Q}(\beta) \quad \text{and} \quad \hat{\beta}(\mathbf{Z}_H) = \arg\min_{\beta\in\mathcal{W}} \hat{Q}_H(\beta),$$

where $\mathcal{W} \subseteq \{v \in \mathbb{R}^J : \|v\|_1 \leq K\}$ and $K > 0$ is a constant. Furthermore, we define $\hat{\Sigma} = T^{-1}\sum_{t=1}^T X_t X_t'$ and $\hat{\mu} = T^{-1}\sum_{t=1}^T X_t Y_t$. Similarly, for $H \subset \{1, ..., T\}$, let $\hat{\Sigma}_H = T^{-1}\sum_{t=1}^T X_{t,H}X_{t,H}'$ and $\hat{\mu} = T^{-1}\sum_{t=1}^T X_{t,H}Y_{t,H}$. Finally, let $\mathcal{H}$ be a set of subsets of $\{1,\ldots,T\}$.

Lemma 10. *Suppose that the following conditions hold:*
(1) with probability at least $1-\gamma_{1,T}$, $\|\hat{\Sigma}_H - \hat{\Sigma}\|_\infty \leq c_T$ and $\|\hat{\mu}_H - \hat{\mu}\|_\infty \leq c_T$ for all $H \in \mathcal{H}$.
(2) with probability at least $1-\gamma_{2,T}$, $\min_{\|v\|_0\leq s} v'\hat{\Sigma}v/\|v\|_2^2 \geq \kappa_1$.

(3) with probability at least $1-\gamma_{3,T}$*,* $\max_{H\in\mathcal{H}}\|\hat{\beta}(\mathbf{Z}_H)\|_0 \leq s/2$ *and* $\|\hat{\beta}(\mathbf{Z})\|_0 \leq s/2$*.*
(4) $P(\max_{1\leq t\leq T}\|X_t\|_\infty \leq \kappa_2)=1$*.*
Let $\hat{\varepsilon}_t = Y_t - X_t'\hat{\beta}(\mathbf{Z})$ *and* $\hat{\varepsilon}_{t,H} = Y_t - X_t'\hat{\beta}(\mathbf{Z}_H)$*. Then we have that*

$$P\left(\max_{H\in\mathcal{H}}\max_{1\leq t\leq T}|\hat{\varepsilon}_t-\hat{\varepsilon}_{t,H}| \leq 2\kappa_2\sqrt{\kappa_1 s c_T K(2K+1)}\right)\geq 1-\gamma_{1,T}-\gamma_{2,T}-\gamma_{3,T}.$$

Lemma 10 provides sufficient conditions for perturbation stability. Inspecting the proof, we notice that the argument does not require the estimator to converge to anything. To the best of our knowledge, this is the first result of this kind. In the conformal prediction literature, one-observation perturbation stability has been considered in Assumption A3 of Lei et al. (2018), who only verify it assuming correct model specification and consistent variable selection. There is also a strand of literature in statistics that considers misspecified models in high dimensions and focuses on the pseudo-true value. For example, for linear models, the pseudo-true value represents the best linear projection and is often assumed to be sparse, making it possible to establish consistency of Lasso to this pseudo-true value, e.g., Bühlmann and van de Geer (2015). We do not make these assumptions. Lemma 10 allows the model to be misspecified and the pseudo-true value may or may not be consistently estimated by constrained Lasso.

Lemma 10 says that when the solution of constrained Lasso is sparse, the stability of $\hat{\Sigma}$ and $\hat{\mu}$ would guarantee the stability of the estimator. When $|H|\asymp\log T_0$ and the observed variables are bounded, we can choose $c_T\asymp T_0^{-1}\log(T_0)$. The sparse eigenvalue condition can typically be verified whenever $s\leq cT$, where $c>0$ is a constant that depends on the eigenvalues of $E\hat{\Sigma}$. Thus, Lemma 10 would guarantee that when $\sup_{H\in\mathcal{H}}|H|\lesssim\log T_0$, we have

$$\max_{H\in\mathcal{H}}\max_{1\leq t\leq T}|\hat{\varepsilon}_t-\hat{\varepsilon}_{t,H}| = O_P(\sqrt{sT_0^{-1}\log T_0}).$$

Therefore, whenever the solutions $\hat{\beta}(\mathbf{Z})$ and $\hat{\beta}(\mathbf{Z}_H)$ are sparse enough with $s=o(T_0/\log(T_0))$, we can expect stability of the estimated residuals. One implication is that since $\|\hat{\beta}(\mathbf{Z})\|_0$ and $\|\hat{\beta}(\mathbf{Z}_H)\|_0$ are clearly bounded above by J, the stability should easily hold for $J\ll T_0/\log(T_0)$.

7 Simulations

This section presents simulation evidence on the finite sample properties of our inference procedure. For concreteness, we focus on the three different methods for estimating counterfactual mean proxies that we are using in our empirical application in Section 8:

difference-in-differences, canonical SC, and constrained Lasso.

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

$$Y_{1t} = \begin{cases} \sum_{j=2}^{J+1} w_j Y_{jt} + u_t & \text{if } t \leq T_0, \\ \alpha_t + \sum_{j=2}^{J+1} w_j Y_{jt} + u_t & \text{if } t > T_0, \end{cases}$$

where $u_t = \rho_u u_{t-1} + v_t$, $v_t \overset{iid}{\sim} N(0, 1 - \rho_u^2)$. Similar to Hahn and Shi (2016), the control outcomes are generated using a factor structure:

$$Y_{jt}^N = \mu_j + \theta_t + \lambda_j F_t + \epsilon_{jt},$$

where $\mu_j = j/J$, $\lambda_j = j/J$, $\theta_t \overset{iid}{\sim} N(0,1)$, $F_t \overset{iid}{\sim} N(0,1)$, and $\epsilon_{jt} = \rho_\epsilon \epsilon_{jt-1} + \xi_{jt}$, $\xi_{jt} \overset{iid}{\sim} N(0, 1 - \rho_\epsilon^2)$. In the simulations, we let $T_* = 1$ and vary ρ_u, ρ_ϵ, T_0, and J. The four DGPs differ with respect to the specification of the weights w.

	Weight specification	**Correctly specified model(s)**
DGP1	$w = \left(\frac{1}{J}, \ldots, \frac{1}{J}\right)'$	DiD, SC, constr. Lasso
DGP2	$w = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, \ldots, 0\right)'$	SC, constr. Lasso
DGP3	$w = -1 \cdot \left(\frac{1}{J}, \ldots, \frac{1}{J}\right)'$	constr. Lasso
DGP4	$w = 2 \cdot \left(\frac{1}{J}, \ldots, \frac{1}{J}\right)'$	–

We consider the problem of testing the null hypothesis of a zero effect:

$$H_0 : \alpha_T = 0.$$

The p-values are computed using the set of moving block permutations $\Pi_\rightarrow$. The nominal size is set equal to $\alpha = 0.1$.

Table 1 shows simulation evidence of the size properties of our method when the data are i.i.d. ($\rho_u = \rho_\epsilon = 0$), which implies exchangeability of the residuals (cf. Lemma 1). The simulation evidence confirms our theoretical results. Our procedure achieves exact size control, irrespectively of whether the method used to estimate P_t^N is correctly specified or not. To study the finite sample performance with dependent data, we set $\rho_u = \rho_\epsilon = 0.6$. Table 2 shows that our method exhibits close-to-correct size, even when the model for P_t^N is misspecified.

We now turn to the power properties of our method. Tables 3 and 4 show finite sample power for a fixed alternative $\alpha_T = 2$ for i.i.d. data ($\rho_u = \rho_\epsilon = 0$) and weakly dependent data ($\rho_u = \rho_\epsilon = 0.6$), respectively. In addition, Figures 2 – 5 displays power curves for a setting where $T_0 = 39$ and $J = 50$ as in our empirical application. Under correct specification, our method exhibits favorable finite sample power properties, irrespective of the specific approach used for estimating P_t^N. However, power can be substantially lower under misspecification. For example, consider the results for DGP3 with i.i.d. data. Power is about three to five times higher when P_t^N is estimated using the correctly specified constrained Lasso estimator than when P_t^N is estimated using the misspecified difference-in-differences or SC estimators. In fact, the power based on the SC estimator is very close to the nominal size for a range of values of the alternative. Thus, while misspecification does not affect size, it may affect power. While the qualitative implications of the results in Tables 3 and 4 are similar, the power of our approach tends to be higher with dependent data.[19]

8 Application: The Impact of Decriminalizing Indoor Prostitution

We revisit the analysis in Cunningham and Shah (2018), who study the causal effect of decriminalizing indoor prostitution on the composition of the sex market, reported rape offenses, and sexually transmitted infections. They exploit that a Rhode Island District Court judge unexpectedly decriminalized indoor sex work in July 2003. Indoor prostitution was eventually re-criminalized in November 2009, but for more than six years Rhode Island was the only US state with decriminalized indoor prostitution and prohibited street prostitution.

We focus on the effect of legalizing indoor prostitution on reported rape offenses and female gonorrhea incidence. Our two outcomes of interest are reported rape rates per 100,000 and log female gonorrhea incidence per 100,000. We use the same data as in Cunningham and Shah (2018).[20] The data on rape offenses come from the Uniform Crime Reports (UCR); the data on gonorrhea cases are from the Center for Disease Control (CDC)'s Gonorrhea Surveillance Program. We refer to Section 3 in Cunningham and Shah (2018) for a detailed description of the data and descriptive statistics. The rape data go back to 1965 such that $T_0 = 39$; the female gonorrhea series date back to 1985 such that $T_0 = 19$. For both outcomes the number of treated periods is $T_1 = 6$. Figure 6 displays the raw data for Rhode Island and the rest of the U.S. states.

[19] We note that, in our simulation setting, power is higher with weakly dependent data even in the "oracle case" where P_t^N is known. The power differences vanish when T_0 grows large.

[20] Following Cunningham and Shah (2018), we smooth the rape series using the moving average of the current and the previous year's rapes.

For our analysis, we work with de-trended data as both series exhibit long run time trends.[21] We compare the results based on three different CSC methods: difference-in-differences, canonical SC, and constrained Lasso. As discussed in more detail in Section 2.3, these methods all specify the counterfactual mean proxy P_t^N as a linear function of the control outcomes:

$$P_t^N = \mu + \sum_{j=2}^{J+1} w_j Y_{jt}.$$

The three methods differ with respect to the restrictions they impose on μ and w. Difference-in-differences leaves μ unrestricted but restricts the weights to be $w_j = 1/J$ same across all control units. Canonical SC imposes that $\mu = 0$ and restricts the weights to be positive and to sum up to one. Constrained Lasso does not impose any restrictions on μ, but restricts the weights to lie in a ℓ_1-ball with radius one. Constrained Lasso thus nests both difference-in-differences and SC. Following Cunningham and Shah (2018), the set of potential control units includes all other U.S. states.

Before turning to the main results, we use the placebo specification tests described in Section 4.3 to assess the plausibility of the underlying assumptions. Specifically, based on the pre-treatment data, we test

$$H_0 : \alpha_{2003-\tau+1} = \cdots = \alpha_{2003} = 0,$$

for $\tau \in \{1, 2, 3\}$. Table 5 shows that we cannot reject the null hypothesis at the conventional significance levels for both outcomes, all three methods, and both types of permutations. This provides evidence in favor of our model specifications and the validity of the maintained assumptions. Figures 7 and 8 provide a graphical illustration of these tests by plotting histograms of the residuals in the pre-treatment period.

Table 6 reports p-values from testing the null hypothesis of a zero effect:

$$H_0 : \alpha_{2004} = \alpha_{2005} = \cdots = \alpha_{2009} = 0. \tag{22}$$

The null hypothesis (22) can rejected at the 5% level for both outcomes, both permutation schemes, and all three methods. Figures 9 and 10 display pointwise 90% confidence intervals. We find similar results for all three methods, which suggest that, while the effect was not or only marginally significant in the first year, legalizing indoor prostitution significantly decreased both rape rates and the incidence of female gonorrhea thereafter, corroborating the findings by Cunningham and Shah (2018).

[21] Specifically, we de-trend all the state time series separately using a quadratic time trend which is estimated based on pre-treatment data for Rhode Island and on all periods for the control states.

9 Conclusion

This paper proposes new inference procedures for counterfactual and synthetic control methods for evaluating policy effects. Our procedures work in conjunction with a great variety of powerful methods for estimating the counterfactual mean outcome in the absence of a policy intervention. The proposed approach has a double justification in that the inference result is exact under strong assumptions on the data, and is approximately exact under weak assumptions on the data and the specific approach used for estimating the counterfactual mean proxies. The proposed approach demonstrates an excellent performance in simulation experiments, and is taken to a data application, where we reevaluate the causal effect of decriminalizing indoor prostitution on rape rates and sexually transmitted infections.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

Bühlmann, P. and van de Geer, S. (2015). High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics*, 9(1):1449–1473.

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.

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

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

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

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., Shao, Q.-M., Wu, W. B., and Xu, L. (2016). Self-normalized cramér-type moderate deviations under dependence. *The Annals of Statistics*, 44(4):1593–1617.

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., Wuthrich, K., and Zhu, Y. (2018). Exact and robust conformal inference methods for predictive machine learning with dependent data.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507.

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.

Dehejia, R. H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84(1):151–161.

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

Ferman, B. and Pinto, C. (2017a). Inference in differences-in-differences with few treated groups and heteroskedasticity.

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

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

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. (2016). Synthetic control and inference. Mimeo.

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

Hansen, C. and Liao, Y. (2016). The factor-lasso and k-step bootstrap approach for inference in high-dimensional economic applications.

Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654.

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.

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

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.

Kosorok, M. R. (2007). *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media.

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*, pages 1–18.

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. (2017). Statistical inference for average treatment effects estimated by synthetic control methods.

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

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.

McCarthy, C. A. (1967). Cp. *Israel Journal of Mathematics*, 5(4):249–271.

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.

Peña, V. H., Lai, T. L., and Shao, Q.-M. (2008). *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media.

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.

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.

Rio, E. (2017). *Asymptotic Theory of Weakly Dependent Random Processes*. Springer.

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.

Romano, J. P. and Shaikh, A. M. (2012). On the uniform asymptotic validity of subsampling and the bootstrap. *The Annals of Statistics*, 40(6):2798–2822.

Rotfeld, S. Y. (1969). The singular numbers of the sum of completely continuous operators. In *Spectral Theory*, pages 73–78. Springer.

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.

Stock, J. and Watson, M. (2016). Factor models and structural vector autoregressions in macroeconomics.

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.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

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

Vovk, V., Nouretdinov, I., and Gammerman, 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.

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

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

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.

Appendix to "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls"

Notations

We introduce some additional notations that will be used in the rest of the paper. For $a, b \in \mathbb{R}$, $a \vee b = \max\{a, b\}$. For two positive sequences a_n, b_n, we use $a_n \ll b_n$ to denote $a_n = o(b_n)$. We use $\Phi(\cdot)$ to denote the cumulative distribution function of the standard normal distribution. Unless stated otherwise, $\|\cdot\|$ denotes the Euclidean norm for vectors or the spectral norm for matrices. We use $\overset{d}{=}$ to denote equal in distribution.

A Proofs

A.1 Proof of Theorem 1

We start with some preliminary definitions and observations. Let $\{S^{(j)}(\hat{u})\}_{j=1}^{n}$ denoted the non-decreasing rearrangement of $\{S(\hat{u}_\pi) : \pi \in \Pi\}$, where $n = |\Pi|$. Call these randomization quantiles. The p-value is defined as

$$\hat{p} = \frac{1}{n} \sum_{\pi \in \Pi} \mathbf{1}(S(\hat{u}_\pi) \geq S(\hat{u})).$$

Observe that

$$\mathbf{1}(\hat{p} \leq \alpha) = \mathbf{1}(S(\hat{u}) > S^{(k)}(\hat{u})),$$

where $k = k(\alpha) = n - \lfloor n\alpha \rfloor = \lceil n(1-\alpha) \rceil$.

The first part of the theorem follows because the Π considered all form a group in the sense that $\Pi\pi = \Pi$ for all $\pi \in \Pi$. The proof uses standard arguments (e.g., Romano, 1990). Because Π forms a group, the randomization quantiles are invariant surely,

$$S^{(k(\alpha))}(\hat{u}_\pi) = S^{(k(\alpha))}(\hat{u}), \text{ for all } \pi \in \Pi.$$

Therefore,

$$\sum_{\pi \in \Pi} \mathbf{1}(S(\hat{u}_\pi) > S^{(k(\alpha))}(\hat{u}_\pi)) = \sum_{\pi \in \Pi} \mathbf{1}(S(\hat{u}_\pi) > S^{(k(\alpha))}(\hat{u})) \leq \alpha n.$$

Since $\mathbf{1}(S(\hat{u}) > S^{(k(\alpha))}(\hat{u}))$ is equal in law to $\mathbf{1}(S(\hat{u}_\pi) > S^{(k(\alpha))}(\hat{u}_\pi))$ for any $\pi \in \Pi$ by exchangeability, we have that

$$\alpha \geq E \sum_{\pi \in \Pi} \mathbf{1}(S(\hat{u}_\pi) > S^{(k(\alpha))}(\hat{u}_\pi))/n = E\mathbf{1}(S(\hat{u}) > S^{(k(\alpha))}(\hat{u})) = E\mathbf{1}(\hat{p} \leq \alpha).$$

For the second part, note that because the joint distribution of $\{S(\hat{u}_\pi)\}_{\pi \in \Pi}$ is continuous, there are no ties with probability one. Therefore,

$$\sum_{\pi\in\Pi} \mathbf{1}(S(\hat{u}_\pi) \leq S^{(k(\alpha))}(\hat{u})) = k(\alpha) \leq n(1-\alpha) + 1$$

Because

$$\sum_{\pi\in\Pi} \mathbf{1}(S(\hat{u}_\pi) \leq S^{(k(\alpha))}(\hat{u})) + \sum_{\pi\in\Pi} \mathbf{1}(S(\hat{u}_\pi) > S^{(k(\alpha))}(\hat{u})) = n,$$

we have that

$$\sum_{\pi\in\Pi} \mathbf{1}(S(\hat{u}_\pi) > S^{(k(\alpha))}(\hat{u})) \geq n\alpha - 1.$$

The result now follows by similar arguments as in the first part.

A.2 Proof of Theorem 2

The proof proceeds by verifying the high-level conditions in the following lemma.

Lemma 11 (Approximate Validity under High-Level Conditions). [22] *Assume that the number of randomizations becomes large, $n = |\Pi| \to \infty$ (in examples above, this is caused by $T \to \infty$). Let $\{\delta_{1n}, \delta_{2n}, \gamma_{1n}, \gamma_{2n}\}$ be sequences of numbers converging to zero, and assume the following conditions.*

(E) With probability $1 - \gamma_{1n}$: the randomization distribution

$$\tilde{F}(x) := \frac{1}{n} \sum_{\pi\in\Pi} \mathbf{1}\{S(u_\pi) < x\},$$

is approximately ergodic for $F(x) = P\left(S(u) < x\right)$, namely

$$\sup_{x\in\mathbb{R}} \left|\tilde{F}(x) - F\left(x\right)\right| \leq \delta_{1n},$$

(A) With probability $1 - \gamma_{2n}$, estimation errors are small:

(1) the mean squared error is small, $n^{-1} \sum_{\pi\in\Pi} \left[S(\hat{u}_\pi) - S(u_\pi)\right]^2 \leq \delta_{2n}^2$;

(2) the pointwise error at $\pi = \text{Identity}$ is small, $|S(\hat{u}) - S(u)| \leq \delta_{2n}$;

(3) The pdf of $S(u)$ is bounded above by a constant D.

Suppose in addition that the null hypothesis is true. Then, the approximate conformal p-value obeys for any $\alpha \in (0, 1)$

$$\left|P\left(\hat{p} \leq \alpha\right) - \alpha\right| \leq 6\delta_{1n} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}}) + \gamma_{1n} + \gamma_{2n}.$$

[22] In Chernozhukov et al. (2018) we use a version of this lemma, which relies on permuting the data instead of permuting the residuals, to derive performance guarantees for prediction intervals obtained using classical conformal prediction methods with weakly dependent data.

With this result at hand, the proof of the theorem is a consequence of following four lemmas, which verify the approximate ergodicity conditions (E) and conditions on the estimation error (A) of Lemma 11. Putting the bounds together and optimizing the error yields the result of the theorem.

The following lemma verifies approximate ergodicity (E) (which allows for large T_*) for the case of moving block permutations.

Lemma 12 (Mixing Implies Approximate Ergodicity)**.** *Let* Π *be the moving block permutations. Suppose that* $\{u_t\}_{t=1}^T$ *is stationary and strong mixing. Assume the following conditions: (1)* $\sum_{k=1}^{\infty} \alpha_{mixing}(k)$ *is bounded by a constant* M*, (2)* $T_0 \geq T_* + 2$*, and (3)* $S(u)$ *has bounded pdf. Then there exists a constant* $M' > 0$ *depending only on* M *such that for any* $\delta_{1n} > 0$*,*

$$P\left(\sup_{x\in\mathbb{R}} \left|\tilde{F}(x) - F(x)\right| \leq \delta_{1n}\right) \geq 1 - \gamma_T,$$

where $\gamma_T = \left(M'\sqrt{\frac{T_*}{T_0}}\log T_0 + \frac{T_*+1}{T_0+T_*}\right)/\delta_{1n}$*.*

The following lemma verifies approximate ergodicity (E) (which allows for large T_*) for the case of i.i.d. permutations.

Lemma 13 (Approximate Ergodicity under i.i.d. Permutations)**.** *Let* Π *be the set of all permutations. Suppose that* $\{u_t\}_{t=1}^T$ *is i.i.d. Assume that* $S(u)$ *only depends on the last* T_* *entries of* u*. If* $T_0 \geq T_* + 2$*, then*

$$P\left(\sup_{x\in\mathbb{R}} \left|\tilde{F}(x) - F(x)\right| \leq \delta_{1n}\right) \geq 1 - \gamma_T,$$

where $\gamma_T = \sqrt{\pi/(2\lfloor T/T_*\rfloor)}/\delta_{1n}$*.*

The following lemma verifies the condition on the estimation error (A) for moving block permutations.

Lemma 14 (Bounds on Estimation Errors under Moving Block Permutations)**.** *Consider moving block permutations* Π*. Let* T_* *be fixed. Suppose that for some constant* $Q > 0$*,* $|S(u) - S(v)| \leq Q\|D_{T_*}(u-v)\|_2$ *for any* $u, v \in \mathbb{R}^T$ *and*$D_{T_*} := \text{Blockdiag}(0_{T_*}, I_{T_*})$*. Then Condition (A) (1)-(2) is satisfied if there exist sequences* $\gamma_T, \delta_{2n} = o(1)$ *such that with probability at least* $1 - \gamma_T$*,*

$$\|\hat{P}^N - P^N\|_2/\sqrt{T} \leq \delta_{2n} \textit{ and } |\hat{P}_t^N - P_t| \leq \delta_{2n} \textit{ for } T_0 + 1 \leq t \leq T.$$

The following lemma verifies the condition on the estimation error (A) for i.i.d. permutations.

Lemma 15 (Bounds on Estimation Errors under i.i.d. Permutations)**.** *Consider the set of all permutations* Π*. Let* T_* *be fixed. Suppose that for some constant* $Q > 0$*,* $|S(u) - S(v)| \leq Q\|D_{T_*}(u-v)\|_2$ *for any* $u, v \in \mathbb{R}^T$ *and* $D_{T_*} := \text{Blockdiag}(0, I_{T_*})$*. Then Condition (A) (1)-(2) is satisfied if there exist sequences* $\gamma_T, \delta_{2n} = o(1)$ *such that with probability at least* $1 - \gamma_T$*,*

$$\|\hat{P}^N - P^N\|_2/\sqrt{T} \leq \delta_{2n} \textit{ and } |\hat{P}_t^N - P_t| \leq \delta_{2n} \textit{ for } T_0 + 1 \leq t \leq T.$$

Now we conclude the proof of Theorem 2.

For the moving block permutations, let $\delta_{1n} = (T_*/T_0)^{1/4}$. Then we apply Lemma 11 together with Lemmas 12 and 14, obtaining

$$\begin{aligned}
|P(\hat{p} \leq \alpha) - \alpha| &\leq 6\delta_{1n} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}}) + \gamma_{1n} + \gamma_{2n} \\
&\leq 6\delta_{1n} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}}) + \left(M'\sqrt{\frac{T_*}{T_0}}\log T_0 + \frac{T_*+1}{T_0+T_*} \right)/\delta_{1n} + \gamma_{2n} \\
&\leq 6(T_*/T_0)^{1/4} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}}) \\
&\quad + \left(M'\sqrt{\frac{T_*}{T_0}}\log T_0 + \frac{T_*+1}{T_0+T_*} \right)(T_*/T_0)^{-1/4} + \gamma_{2n}.
\end{aligned}$$

The final result for moving block permutations follows by straight-forward computations and the observations that $\delta_{2n} = O(\sqrt{\delta_{2n}})$ (due to $\delta_{2n} = o(1)$).

For i.i.d permutations, we also use $\delta_{1n} = (T_*/T_0)^{1/4}$. Then we apply Lemma 11 together with Lemmas 13 and 15, obtaining

$$\begin{aligned}
|P(\hat{p} \leq \alpha) - \alpha| &\leq 6\delta_{1n} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}}) + \gamma_{1n} + \gamma_{2n} \\
&\leq 6\delta_{1n} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}}) + \sqrt{2\pi/\lfloor T/T_* \rfloor}/\delta_{1n} + \gamma_{2n} \\
&\leq 6(T_*/T_0)^{1/4} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}}) + \sqrt{2\pi/\lfloor T/T_* \rfloor}(T_*/T_0)^{-1/4} + \gamma_{2n} \\
&\lesssim (T_*/T_0)^{1/4} + \delta_{2n} + \sqrt{\delta_{2n}} + \gamma_{2n}.
\end{aligned}$$

This completes the proof for i.i.d. permutations.

A.2.1 Proof of Lemma 11

The proof proceeds in two steps.

Step 1: We bound the difference between the p-value and the oracle p-value, $\hat{F}(S(\hat{u})) - F(S(u))$.

Let $\mathcal{M}$ be the event that the conditions (A) and (E) hold. By assumption,

$$P(\mathcal{M}) \geq 1 - \gamma_{1n} - \gamma_{2n}. \tag{23}$$

Notice that on the event $\mathcal{M}$,

$$\begin{aligned}
\left|\hat{F}(S(\hat{u})) - F(S(u))\right| &\leq \left|\hat{F}(S(\hat{u})) - F(S(\hat{u}))\right| + |F(S(\hat{u})) - F(S(u))| \\
&\overset{\text{(i)}}{\leq} \sup_{x\in\mathbb{R}} \left|\hat{F}(x) - F(x)\right| + D\,|S(\hat{u}) - S(u)| \\
&\leq \sup_{x\in\mathbb{R}} \left|\hat{F}(x) - \tilde{F}(x)\right| + \sup_{x\in\mathbb{R}} \left|\tilde{F}(x) - F(x)\right| + D\,|S(\hat{u}) - S(u)| \\
&\leq \sup_{x\in\mathbb{R}} \left|\hat{F}(x) - \tilde{F}(x)\right| + \delta_{1n} + D\,|S(\hat{u}) - S(u)|
\end{aligned}$$

$$\leq \sup_{x\in\mathbb{R}} \left|\hat{F}(x) - \tilde{F}(x)\right| + \delta_{1n} + D\delta_{2n}, \tag{24}$$

where (i) holds by the fact that the bounded pdf of $S(u)$ implies Lipschitz property for F.

Let $A = \left\{\pi \in \Pi : \ |S(\hat{u}_\pi) - S(u_\pi)| \geq \sqrt{\delta_{2n}}\right\}$. Observe that on the event $\mathcal{M}$, by Chebyshev inequality

$$|A|\delta_{2n} \leq \sum_{\pi\in\Pi} (S(\hat{u}_\pi) - S(u_\pi))^2 \leq n\delta_{2n}^2$$

and thus $|A|/n \leq \delta_{2n}$. Also observe that on the event $\mathcal{M}$, for any $x \in \mathbb{R}$,

$$\begin{aligned}
&\left|\hat{F}(x) - \tilde{F}(x)\right| \\
&\leq \frac{1}{n}\sum_{\pi\in A} \left|\mathbf{1}\left\{S(\hat{u}_\pi) < x\right\} - \mathbf{1}\left\{S(u_\pi) < x\right\}\right| + \frac{1}{n}\sum_{\pi\in(\Pi\backslash A)} \left|\mathbf{1}\left\{S(\hat{u}_\pi) < x\right\} - \mathbf{1}\left\{S(u_\pi) < x\right\}\right| \\
&\overset{\text{(i)}}{\leq} 2\frac{|A|}{n} + \frac{1}{n}\sum_{\pi\in(\Pi\backslash A)} \mathbf{1}\left\{|S(u_\pi) - x| \leq \sqrt{\delta_{2n}}\right\} \leq 2\frac{|A|}{n} + \frac{1}{n}\sum_{\pi\in\Pi} \mathbf{1}\left\{|S(u_\pi) - x| \leq \sqrt{\delta_{2n}}\right\} \\
&\leq 2\frac{|A|}{n} + P\left(|S(u) - x| \leq \sqrt{\delta_{2n}}\right) \\
&\quad + \sup_{z\in\mathbb{R}} \left|\frac{1}{n}\sum_{\pi\in\Pi} \mathbf{1}\left\{|S(u_\pi) - z| \leq \sqrt{\delta_{2n}}\right\} - P\left(|S(u) - z| \leq \sqrt{\delta_{2n}}\right)\right| \\
&= 2\frac{|A|}{n} + P\left(|S(u) - x| \leq \sqrt{\delta_{2n}}\right) \\
&\quad + \sup_{x\in\mathbb{R}} \left|\left[\tilde{F}\left(z + \sqrt{\delta_{2n}}\right) - \tilde{F}\left(z - \sqrt{\delta_{2n}}\right)\right] - \left[F\left(z + \sqrt{\delta_{2n}}\right) - F\left(z - \sqrt{\delta_{2n}}\right)\right]\right| \\
&\leq 2\frac{|A|}{n} + P\left(|S(u) - x| \leq \sqrt{\delta_{2n}}\right) \\
&\quad + 2\sup_{z\in\mathbb{R}} \left|\tilde{F}(z) - F\left(z\right)\right| \\
&\overset{\text{(ii)}}{\leq} 2\frac{|A|}{n} + 2D\sqrt{\delta_{2n}} + 2\delta_{1n} \overset{\text{(iii)}}{\leq} 2\delta_{1n} + 2\delta_{2n} + 2D\sqrt{\delta_{2n}},
\end{aligned} \tag{25}$$

where (i) follows by the boundedness of indicator functions and the elementary inequality of $|\mathbf{1}\{S(\hat{u}_\pi) < x\} - \mathbf{1}\{S(u_\pi) < x\}| \leq \mathbf{1}\{|S(u_\pi) - x| \leq |S(\hat{u}_\pi) - S(u_\pi)|\}$, (ii) follows by the bounded pdf of $S(u)$ and (iii) follows by $|A|/n \leq \delta_{2n}$. Since the above display holds for each $x \in \mathbb{R}$, it follows that on the event $\mathcal{M}$,

$$\sup_{x\in\mathbb{R}} \left|\hat{F}(x) - \tilde{F}(x)\right| \leq 2\delta_{1n} + 2\delta_{2n} + 2D\sqrt{\delta_{2n}}. \tag{26}$$

We combine (24) and (26) and obtain that on the event $\mathcal{M}$,

$$\left|\hat{F}(S(\hat{u})) - F(S(u))\right| \leq 3\delta_{1n} + 2\delta_{2n} + D(\delta_{2n} + 2\sqrt{\delta_{2n}}). \tag{27}$$

Step 2: Here we derive the desired result. Notice that

$$
\begin{aligned}
& \left|P\left(1-\hat{F}(S(\hat{u})) \leq \alpha\right)-\alpha\right| \\
& =\left|E\left(\mathbf{1}\left\{1-\hat{F}(S(\hat{u})) \leq \alpha\right\}-\mathbf{1}\{1-F(S(u)) \leq \alpha\}\right)\right| \\
& \leq E\left|\mathbf{1}\left\{1-\hat{F}(S(\hat{u})) \leq \alpha\right\}-\mathbf{1}\{1-F(S(u)) \leq \alpha\}\right| \\
& \overset{\text{(i)}}{\leq} P\left(|F(S(u))-1+\alpha| \leq\left|\hat{F}(S(\hat{u}))-F(S(u))\right|\right) \\
& \leq P\left(|F(S(u))-1+\alpha| \leq\left|\hat{F}(S(\hat{u}))-F(S(u))\right| \text{ and } \mathcal{M}\right)+P(\mathcal{M}^c) \\
& \overset{\text{(ii)}}{\leq} P\left(|F(S(u))-1+\alpha| \leq 3\delta_{1n}+2\delta_{2n}+D(\delta_{2n}+2\sqrt{\delta_{2n}})\right)+P\left(\mathcal{M}^c\right) \\
& \overset{\text{(iii)}}{\leq} 6\delta_{1n}+4\delta_{2n}+2D(\delta_{2n}+2\sqrt{\delta_{2n}})+\gamma_{1n}+\gamma_{2n},
\end{aligned}
$$

where (i) follows by the elementary inequality $|\mathbf{1}\{1-\hat{F}(S(\hat{u})) \leq \alpha\}-\mathbf{1}\{1-F(S(u)) \leq \alpha\}| \leq \mathbf{1}\{|F(S(u))-1+\alpha| \leq |\hat{F}(S(\hat{u}))-F(S(u))|\}$, (ii) follows by (27), (iii) follows by the fact that $F(S(u))$ has the uniform distribution on $(0,1)$ and hence has pdf equal to 1, and by (23). The proof is complete.

A.2.2 Proof of Lemma 12

We define

$$
s_t=\begin{cases}(\sum_{s=t}^{t+T_*-1}|u_s|^q)^{1/q} & \text{if } 1 \leq t \leq T_0 \\ (\sum_{s=t}^{T}|u_s|^q+\sum_{s=1}^{t-T_0-1}|u_s|^q)^{1/q} & \text{otherwise.}\end{cases}
$$

It is straight-forward to verify that

$$
\{S(u_\pi):\ \pi \in \Pi\}=\{s_t:\ 1 \leq t \leq T\}.
$$

Let $\tilde{\alpha}_{\text{mixing}}$ be the strong-mixing coefficient for $\{s_t\}_{t=1}^{T_0}$. Notice that $\{s_t\}_{t=1}^{T_0}$ is stationary (although $\{s_t\}_{t=1}^{T}$ is clearly not). Let $\check{F}(x)=T_0^{-1}\sum_{t=1}^{T_0}\mathbf{1}\{s_t \leq x\}$. The bounded pdf of $S(u)$ implies the continuity of $F(\cdot)$. It follows, by Proposition 7.1 of Rio (2017), that

$$
E\left(\sup_{x\in\mathbb{R}}\left|\check{F}(x)-F(x)\right|^2\right) \leq \frac{1}{T_0}\left(1+4\sum_{k=0}^{T_0-1}\tilde{\alpha}_{\text{mixing}}(t)\right)\left(3+\frac{\log T_0}{2\log 2}\right)^2. \tag{28}
$$

Notice that $\tilde{\alpha}_{\text{mixing}}(t) \leq 2$ and that $\tilde{\alpha}_{\text{mixing}}(t) \leq \alpha_{\text{mixing}}\left(\max\{t-T_*,0\}\right)$ so that

$$
\begin{aligned}
\sum_{k=0}^{T_0-1}\tilde{\alpha}_{\text{mixing}}(t) &= \sum_{k=0}^{T_*}\tilde{\alpha}_{\text{mixing}}(t)+\sum_{k=T_*+1}^{T_0-1}\tilde{\alpha}_{\text{mixing}}(t) \leq 2(T_*+1)+\sum_{k=1}^{T_0-T_*-1}\alpha_{\text{mixing}}(k) \\
&\leq 2(T_*+1)+\sum_{k=1}^{\infty}\alpha_{\text{mixing}}(k).
\end{aligned}
$$

Since $\sum_{k=1}^{\infty} \alpha_{\text{mixing}}(k)$ is bounded by M, it follows by (28) that

$$E\left(\sup_{x\in\mathbb{R}} \left|\check{F}(x) - F(x)\right|^2\right) \le B_T := \frac{1+4(2(T_*+1)+M)}{T_0}\left(3+\frac{\log T_0}{2\log 2}\right)^2.$$

By Liapunov's inequality,

$$E\left(\sup_{x\in\mathbb{R}} \left|\check{F}(x) - F(x)\right|\right) \le \sqrt{E\left(\sup_{x\in\mathbb{R}} \left|\check{F}(x) - F(x)\right|^2\right)} \le \sqrt{B_T}.$$

Since $(T_0+T_*)\tilde{F}(x) - T_0\check{F}(x) = \sum_{t=T_0+1}^{T_0+T_*} \mathbf{1}\{s_t \le x\}$, it follows that

$$\begin{aligned}\sup_{x\in\mathbb{R}} \left|\tilde{F}(x) - \check{F}(x)\right| &= \sup_{x\in\mathbb{R}} \left|\left(\frac{T_0}{T_0+T_*}\check{F}(x) + \frac{1}{T_0+T_*}\sum_{t=T_0+1}^{T_0+T_*} \mathbf{1}\{s_t \le x\}\right) - \check{F}(x)\right| \\ &= \sup_{x\in\mathbb{R}} \left|\frac{1}{T_0+T_*}\check{F}(x) + \frac{1}{T_0+T_*}\sum_{t=T_0+1}^{T_0+T_*} \mathbf{1}\{s_t \le x\}\right| \le \frac{T_*+1}{T_0+T_*},\end{aligned}$$

where the last inequality follows by $\sup_{x\in\mathbb{R}} |\check{F}(x)| \le 1$ and the boundedness of the indicator function. Combining the above two displays, we obtain that

$$E\left(\sup_{x\in\mathbb{R}} \left|\tilde{F}(x) - F(x)\right|\right) \le \sqrt{B_T} + \frac{T_*+1}{T_0+T_*}.$$

The desired result follows by Markov's inequality.

A.2.3 Proof of Lemma 13

The proof follows by an argument given by Romano and Shaikh (2012) for subsampling. We give a complete argument for our setting here for clarity and completeness.

Recall that Π is the set of all bijections π on $\{1, ..., T\}$. Let $k_T = \lfloor T/T_* \rfloor$. Define the blocks of indices

$$b_i = (T - iT_* + 1, T - iT_* + 2, ..., T - iT_* + T_*) \in \mathbb{R}^{T_*}, \qquad i = 1, ..., k_T$$

Since $S(u)$ only depends on u_{b_1}, the last T_* entries of u, we can define

$$Q(x; u_{b_1}) = \mathbf{1}\{S(u) \le x\} - F(x).$$

Therefore,

$$\tilde{F}(x) - F(x) = \frac{1}{|\Pi|}\sum_{\pi\in\Pi} Q(u_{\pi(b_1)}; x).$$

Define $\pi(b_i) := \pi_{|b_i}(b_i)$ to mean the restriction of the permutation map $\pi : \{1, \ldots T\} \to \{1, \ldots T\}$ to the domain b_i.

Notice that for $1 \le i \le k_T$, the value of $\sum_{\pi\in\Pi} Q(u_{\pi(b_i)}; x)$ does not depend on i. It follows that

$$\tilde{F}(x) - F(x) = \frac{1}{|\Pi|}\sum_{\pi\in\Pi} Q(u_{\pi(b_1)};x) = \frac{1}{k_T}\sum_{i=1}^{k_T}\left(\frac{1}{|\Pi|}\sum_{\pi\in\Pi} Q(u_{\pi(b_i)};x)\right)$$
$$= \frac{1}{|\Pi|}\sum_{\pi\in\Pi}\left[\frac{1}{k_T}\sum_{i=1}^{k_T} Q(u_{\pi(b_i)};x)\right].$$

Hence by Jensen's inequality

$$E\left(\sup_{x\in\mathbb{R}}\left|\tilde{F}(x) - F(x)\right|\right) \le \frac{1}{|\Pi|}\sum_{\pi\in\Pi} E\left(\sup_{x\in\mathbb{R}}\left|\frac{1}{k_T}\sum_{i=1}^{k_T} Q(u_{\pi(b_i)};x)\right|\right).$$

To compute the above expectation, we observe that for any $\pi \in \Pi$,

$$E\left(\sup_{x\in\mathbb{R}}\left|\frac{1}{k_T}\sum_{i=1}^{k_T} Q(u_{\pi(b_i)};x)\right|\right) = \int_0^1 P\left(\sup_{x\in\mathbb{R}}\left|\frac{1}{k_T}\sum_{i=1}^{k_T} Q(u_{\pi(b_i)};x)\right| > z\right) dz$$
$$\le \int_0^1 2\exp\left(-2k_T z^2\right) dz < \int_0^\infty 2\exp\left(-2k_T z^2\right) dz = \sqrt{\pi/(2k_T)},$$

where the first inequality follows by the Dvoretsky-Kiefer-Wolfwitz inequality (e.g., Theorem 11.6 in Kosorok (2007)) and the fact that for any $\pi \in \Pi$, $\{Q(u_{\pi(b_i)};x)\}_{i=1}^{k_T}$ is a sequence of i.i.d random variables (since π is a bijection and $\{b_i\}_{i=1}^{k_T}$ are disjoint blocks of indices); the last equality follows from the properties of the normal density. Therefore, the above two display imply that

$$E\left(\sup_{x\in\mathbb{R}}\left|\tilde{F}(x) - F(x)\right|\right) \le \sqrt{\pi/(2k_T)}.$$

The desired result follows by Markov's inequality.

A.2.4 Proof of Lemma 14

Due to the Lipschitz property of $S(\cdot)$, we have

$$\sum_{\pi\in\Pi}\left[S(\hat{u}_\pi) - S(u_\pi)\right]^2 \le Q\sum_{\pi\in\Pi}\|D_{T_*}(\hat{u}_\pi - u_\pi)\|_2^2 = Q\sum_{\pi\in\Pi}\sum_{t=T_0+1}^{T_0+T_*}\left(\hat{u}_{\pi(t)} - u_{\pi(t)}\right)^2$$
$$= Q\sum_{t=T_0+1}^{T_0+T_*}\sum_{\pi\in\Pi}\left(\hat{u}_{\pi(t)} - u_{\pi(t)}\right)^2 = QT_*\|\hat{u} - u\|_2^2 = QT_*\|\hat{P}^N - P^N\|^2$$

where the penultimate equality follows by the observation that for moving block permutation Π,

$$\sum_{\pi\in\Pi}\left(\hat{u}_{\pi(t)} - u_{\pi(t)}\right)^2 = \|\hat{u} - u\|_2^2.$$

Hence condition (A) (1) follows with a rescaled value of δ_n. Condition (A) (2) holds by the Lipschitz property of $S(\cdot)$:

$$|S(\hat{u}) - S(u)| \leq Q\|D_{T_*}(\hat{u} - u)\|_2 \leq Q\sqrt{\sum_{t=T_0+1}^{T_0+T_*} (\hat{u}_t - u_t)^2}$$

Hence, Condition (A) (2) follows since $\|\hat{P}_t^N - P_t^N\| = |\hat{u}_t - u_t| \leq \delta_n$ for $T_0 + 1 \leq t \leq T$ with high probability. The proof is complete.

A.2.5 Proof of Lemma 15

For $t, s \in \{1, ..., T\}$, we define $A_{t,s} = \{\pi \in \Pi : \pi(t) = s\}$. Recall that Π is the set of all bijections on $\{1, ..., T\}$. Thus, $|A_{t,s}| = (T-1)!$. It follows that for any $t \in \{1, ..., T\}$,

$$\begin{aligned}\sum_{\pi\in\Pi} \left(\hat{u}_{\pi(t)} - u_{\pi(t)}\right)^2 &= \sum_{s=1}^{T} \sum_{\pi\in A_{t,s}} \left(\hat{u}_{\pi(t)} - u_{\pi(t)}\right)^2 \\ &= \sum_{s=1}^{T} \sum_{\pi\in A_{t,s}} \left(\hat{u}_s - u_s\right)^2 = \sum_{s=1}^{T} |A_{t,s}| \left(\hat{u}_s - u_s\right)^2 = (T-1)! \times \|\hat{u} - u\|_2^2.\end{aligned} \tag{29}$$

Due to the Lipschitz property of $S(\cdot)$, we have that

$$\begin{aligned}&\frac{1}{|\Pi|}\sum_{\pi\in\Pi} [S(\hat{u}_\pi) - S(u_\pi)]^2 \leq \frac{Q}{|\Pi|}\sum_{\pi\in\Pi} \|D_{T_*}(\hat{u}_\pi - u_\pi)\|_2^2 = \frac{Q}{|\Pi|}\sum_{\pi\in\Pi}\sum_{t=T_0+1}^{T_0+T_*} \left(\hat{u}_{\pi(t)} - u_{\pi(t)}\right)^2 \\ &= \frac{Q}{|\Pi|}\sum_{t=T_0+1}^{T_0+T_*}\sum_{\pi\in\Pi} \left(\hat{u}_{\pi(t)} - u_{\pi(t)}\right)^2 = \frac{Q}{|\Pi|} T_*(T-1)! \times \|\hat{u} - u\|_2^2 = QT^{-1}T_*\|\hat{u} - u\|_2^2,\end{aligned}$$

where the penultimate equality follows by (29) and the last equality follows by $|\Pi| = T!$. Thus, part 1 of Condition (A) follows since T_* is fixed.

To see part 2 of Condition (A), notice that the Lipschitz property of $S(\cdot)$ implies

$$|S(\hat{u}) - S(u)| \leq Q\|D_{T_*}(\hat{u} - u)\|_2 \leq Q\sqrt{\sum_{t=T_0+1}^{T_0+T_*} (\hat{u}_t - u_t)^2}.$$

Hence, part 2 of Condition (A) follows since $|\hat{u}_t - u_t| \leq \delta_n$ for $T_0 + 1 \leq t \leq T$ with high probability. The proof is complete.

A.3 Proof of Theorem 3

We first state an auxiliary lemma.

Lemma 16. *Let $\{W_t\}_{t=1}^T$ be a stationary and β-mixing sequence with coefficient $\beta_{\mathrm{mixing}}(\cdot)$. Let $G(x) = P(W_t \leq x)$. Then for any positive integer $1 \leq m \leq T/2$, we have*

$$E\left(\sup_{x\in\mathbb{R}}\left|T^{-1}\sum_{t=1}^{T}[\mathbf{1}\{W_t \leq x\} - G(x)]\right|\right) \leq 2\sqrt{T}\beta_{\mathrm{mixing}}(m) + \sqrt{\pi m/(2T)} + (m-1)/T.$$

Now we prove Theorem 3. In this proof, universal constants refer to constants that depend only on $D_1, D_2, D_3 > 0$. Define $\tilde{F}(x) = R^{-1}\sum_{j=1}^{R}\tilde{F}_j(x)$, where

$$\tilde{F}_j(x) = m^{-1}\sum_{t\in H_j}\mathbf{1}\left\{\phi\left(Z_t, ..., Z_{t+T_*-1}; \hat{\beta}(\mathbf{Z})\right) \leq x\right\}.$$

Define

$$\hat{F}(x) = T^{-1}\left(\sum_{t=1}^{T_0}\mathbf{1}\left\{\phi\left(Z_t, ..., Z_{t+T_*-1}; \hat{\beta}(\mathbf{Z})\right) \leq x\right\} + \sum_{t=T_0+1}^{T_0+T_*}\mathbf{1}\left\{\phi\left(Z_{q(t)}, ..., Z_{q(t+T_*-1)}; \hat{\beta}(\mathbf{Z})\right) \leq x\right\}\right), \tag{30}$$

where $q(t) = t\mathbf{1}\{t \leq T\} + (t-T)\mathbf{1}\{t > T\}$.

The rest of the proof proceeds in 4 steps. The first three steps bound $\sup_{x\in\mathbb{R}}|\hat{F}(x) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)|$, where we recall that $\Psi(x;\beta) = P(\phi(Z_t, ..., Z_{t+T_*-1}; \beta) \leq x)$. The fourth step derives the desired result.

Step 1: bound $\sup_{x\in\mathbb{R}}\left|\tilde{F}_j(x) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\right|$.

Let $A_j = \bigcup_{t\in H_j}\{t, ..., t+T_*-1\}$. Since $k > T_*$, we have that $A_j \subset \widetilde{H}_j$ and $\min_{t\in A_j,\ s\in\widetilde{H}_j^c}|t-s| \geq k - T_* + 1$. This means that $\{Z_t\}_{t\in\widetilde{H}_j^c}$ and $\{Z_t\}_{t\in A_j}$ have a gap of at least $k - T_* + 1$ time periods. By Berbee's coupling (e.g., Lemma 7.1 of Chen et al. (2016)), there exist random elements $\{\bar{Z}_t\}_{t\in A_j}$ (on an enlarged probability space) such that (1) $\{\bar{Z}_t\}_{t\in A_j}$ is independent of $\{Z_t\}_{t\in\widetilde{H}_j^c}$, (2) $\{\bar{Z}_t\}_{t\in A_j} \overset{d}{=} \{Z_t\}_{t\in A_j}$ and (3) $P(\{\bar{Z}_t\}_{t\in A_j} \neq \{Z_t\}_{t\in A_j}) \leq \beta_{\mathrm{mixing}}(k - T_* + 1)$. Since $\{\tilde{Z}_t\}_{t\in\widetilde{H}_j}$ is independent of the data, we can construct $\{\bar{Z}_t\}_{t\in A_j}$ such that it is also independent of $\mathbf{Z}_{\widetilde{H}_j}$.

Define the event

$$\begin{aligned}\mathcal{M}_j = \left\{\{\bar{Z}_t\}_{t\in A_j} = \{Z_t\}_{t\in A_j}\right\}\bigcap\left\{\sup_{x\in\mathbb{R}}\left|\partial\Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)/\partial x\right| \leq \xi_T\right\}\\ \bigcap\left\{\max_{\pi\in\Pi}\left|S\left(\mathbf{Z}^{\pi}, \hat{\beta}(\mathbf{Z})\right) - S\left(\mathbf{Z}^{\pi}, \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\right| \leq \varrho_T(|\widetilde{H}_j|)\right\}\end{aligned}$$

as well as the functions

$$\begin{cases}\check{F}_j(x) = m^{-1}\sum_{t\in H_j}\mathbf{1}\left\{\phi\left(\bar{Z}_t, ..., \bar{Z}_{t+T_*-1}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) \leq x\right\}\\ \dot{F}_j(x) = m^{-1}\sum_{t\in H_j}\mathbf{1}\left\{\phi\left(Z_t, ..., Z_{t+T_*-1}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) \leq x\right\}.\end{cases}$$

By the construction of $\{\bar{Z}_t\}_{t\in A_j}$ and Assumptions 4 and 5, $P(\mathcal{M}_j^c) \leq \beta_{\text{mixing}}(k - T_* + 1) + \gamma_{1,T} + \gamma_{2,T}$.

Notice that conditional on $\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})$, $\phi\left(\bar{Z}_t, ..., \bar{Z}_{t+T_*-1}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)$ is a stationary β-mixing across $t \in H_j$ with mixing coefficient $\tilde{\beta}_{\text{mixing}}(i) \leq \beta_{\text{mixing}}(i - T_* + 1)$ for $i \geq T_*$. Moreover,

$$\Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) = P\left(\phi\left(\bar{Z}_t, ..., \bar{Z}_{t+T_*-1}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) \leq x \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right).$$

Hence, by Lemma 16, we have that for any $m_1 \leq m/2$,

$$E\left(\sup_{x\in\mathbb{R}}\left|\check{F}_j(x) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\right|\right) \leq 2m^{1/2}\beta_{\text{mixing}}(m_1 - T_* + 1) + \sqrt{\pi m_1/(2m)} + (m_1 - 1)/m.$$

We shall choose m_1 later. Observe that on the event $\mathcal{M}_j$, $\check{F}_j(\cdot) = \dot{F}_j(\cdot)$. Therefore,

$$E(a_j) \leq 2m^{1/2}\beta_{\text{mixing}}(m_1 - T_* + 1) + \sqrt{\pi m_1/(2m)} + (m_1 - 1)/m + 2P(\mathcal{M}_j^c), \tag{31}$$

where $a_j = \sup_{x\in\mathbb{R}}\left|\dot{F}_j(x) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\right|$.

Now we bound $\sup_{x\in\mathbb{R}}|\dot{F}_j(x) - \tilde{F}_j(x)|$. Fix an arbitrary $x \in \mathbb{R}$. Observe that on the event $\mathcal{M}_j$,

$$\begin{aligned}
&\left|\tilde{F}_j(x) - \dot{F}_j(x)\right| \\
&= \left|m^{-1}\sum_{t\in H_j}\left(\mathbf{1}\left\{\phi\left(Z_t, ..., Z_{t+T_*-1}; \hat{\beta}(\mathbf{Z})\right) \leq x\right\} - \mathbf{1}\left\{\phi\left(Z_t, ..., Z_{t+T_*-1}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) \leq x\right\}\right)\right| \\
&\leq m^{-1}\sum_{t\in H_j}\left|\mathbf{1}\left\{\phi\left(Z_t, ..., Z_{t+T_*-1}; \hat{\beta}(\mathbf{Z})\right) \leq x\right\} - \mathbf{1}\left\{\phi\left(Z_t, ..., Z_{t+T_*-1}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) \leq x\right\}\right| \\
&\overset{\text{(i)}}{\leq} m^{-1}\sum_{t\in H_j}\mathbf{1}\left\{\left|\phi\left(Z_t, ..., Z_{t+T_*-1}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) - x\right| \leq \varrho_T(|\widetilde{H}_j|)\right\} \\
&= m^{-1}\sum_{t\in H_j}\mathbf{1}\left\{\phi\left(Z_t, ..., Z_{t+T_*-1}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) \leq x + \varrho_T(|\widetilde{H}_j|)\right\} \\
&\qquad - m^{-1}\sum_{t\in H_j}\mathbf{1}\left\{\phi\left(Z_t, ..., Z_{t+T_*-1}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) < x - \varrho_T(|\widetilde{H}_j|)\right\} \\
&< \dot{F}_j\left(x + \varrho_T(|\widetilde{H}_j|)\right) - \dot{F}_j\left(x - 2\varrho_T(|\widetilde{H}_j|)\right) \\
&\leq \Psi\left(x + \varrho_T(|\widetilde{H}_j|); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) - \Psi\left(x - 2\varrho_T(|\widetilde{H}_j|); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) + 2a_j \overset{\text{(ii)}}{\leq} 3\xi_T\varrho_T(|\widetilde{H}_j|) + 2a_j,
\end{aligned}$$

where (i) follows by the elementary inequality $|\mathbf{1}\{x \leq z\} - \mathbf{1}\{y \leq z\}| \leq \mathbf{1}\{|y - z| \leq |x - y|\}$ for any $x, y, z \in \mathbb{R}$ and (ii) follows by the definition of $\mathcal{M}_j$. Since the above bound holds for any $x \in \mathbb{R}$ and $|\widetilde{H}_j| \leq m + 2k$, we have that on the event $\mathcal{M}_j$,

$$\sup_{x\in\mathbb{R}} \left| \tilde{F}_j(x) - \dot{F}_j(x) \right| \leq 3\xi_T \varrho_T (m+2k) + 2a_j.$$

By the definition of a_j, this means that on the event $\mathcal{M}_j$,

$$\sup_{x\in\mathbb{R}} \left| \tilde{F}_j(x) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) \right| \leq 3\xi_T \varrho_T (m+2k) + 3a_j.$$

By (31) and the fact that $\tilde{F}_j(\cdot)$ and $\Psi(\cdot,\cdot)$ take values in $[0,1]$, we have that for a universal constant $C_1 > 0$,

$$\begin{aligned}
&E\left(\sup_{x\in\mathbb{R}} \left| \tilde{F}_j(x) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) \right| \right) \qquad (32)\\
&\leq 3\xi_T \varrho_T (m+2k) + 3E(a_j) + 2P(\mathcal{M}_j^c)\\
&\leq 3\xi_T \varrho_T (m+2k) + 6m^{1/2}\beta_{\text{mixing}}(m_1 - T_* + 1) + 3\sqrt{\pi m_1/(2m)} + 3(m_1-1)/m + 8P(\mathcal{M}_j^c)\\
&\overset{\text{(i)}}{\leq} C_1 \left(\xi_T \varrho_T (m+2k) + m^{1/2}\beta_{\text{mixing}}(m_1 - T_* + 1) + \sqrt{m_1/m} + \beta_{\text{mixing}}(k - T_* + 1) + \gamma_{1,T} + \gamma_{2,T} \right),
\end{aligned}$$

where (i) follows by $P(\mathcal{M}_j^c) \leq \beta_{\text{mixing}}(k - T_* + 1) + \gamma_{1,T} + \gamma_{2,T}$ and $m_1/m \leq \sqrt{m_1/m}$.

Step 2: bound $R^{-1}\sum_{j=1}^{R} \sup_{x\in\mathbb{R}} \left| \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \right|$.

Let $\dot{\mathbf{Z}} = \{\dot{Z}_t\}_{t=1}^T$ satisfy that $\dot{\mathbf{Z}} \overset{d}{=} \mathbf{Z}$ and $\dot{\mathbf{Z}}$ is independent of $(\mathbf{Z}, \{\tilde{Z}_t\}_{t=1}^T)$. Therefore, for any $1 \leq j \leq R$,

$$\begin{aligned}
\Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) &= P\left(\phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j}) \right) \leq x \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j}) \right)\\
&\overset{\text{(i)}}{=} P\left(\phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j}) \right) \leq x \mid \mathbf{Z}, \{\tilde{Z}_t\}_{t=1}^T \right),
\end{aligned}$$

where (i) follows by the fact that $\dot{\mathbf{Z}}$ and $(\mathbf{Z}, \{\tilde{Z}_t\}_{t=1}^T)$ are independent. (This is the identity that $E(f(X; g(Y)) \mid g(Y)) = E(f(X; g(Y)) \mid Y)$ for any measurable functions f and g if X and Y are independent. To see this, simply notice that the distribution of X given $g(Y)$ and the distribution of X given Y are both equal to the unconditional distribution of X.)

Define the event

$$\mathcal{Q}_j = \left\{ \sup_{x\in\mathbb{R}} \left| \partial\Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)/\partial x \right| \leq \xi_T \right\}.$$

Clearly, $P(\mathcal{Q}_j) \geq 1 - \gamma_{2,T}$ by Assumption 5. Therefore, we have that on the event $\mathcal{Q}_j$,

$$\begin{aligned}
&\left| \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \right|\\
&= \left| E\left(\mathbf{1}\left\{ \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j}) \right) \leq x \right\} - \mathbf{1}\left\{ \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R}) \right) \leq x \right\} \mid \mathbf{Z}, \{\tilde{Z}_t\}_{t=1}^T \right) \right|\\
&\leq E\left(\left| \mathbf{1}\left\{ \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j}) \right) \leq x \right\} - \mathbf{1}\left\{ \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R}) \right) \leq x \right\} \right| \mid \mathbf{Z}, \{\tilde{Z}_t\}_{t=1}^T \right)
\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leq} E\Bigg[\mathbf{1}\Bigg\{\left|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)-x\right| \\
&\qquad\qquad \leq \left|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\right|\Bigg\} \mid \mathbf{Z},\{\tilde{Z}_t\}_{t=1}^T\Bigg] \\
&= P\Bigg[\left|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)-x\right| \\
&\qquad\qquad \leq \left|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\right| \mid \mathbf{Z},\{\tilde{Z}_t\}_{t=1}^T\Bigg] \\
&\leq P\left[\left|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)-x\right| \leq 2\varrho_T(|\widetilde{H}_j|) \mid \mathbf{Z},\{\tilde{Z}_t\}_{t=1}^T\right] \\
&\quad + P\left[\left|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\right| > 2\varrho_T(|\widetilde{H}_j|) \mid \mathbf{Z},\{\tilde{Z}_t\}_{t=1}^T\right] \\
&= \Psi\left(x+2\varrho_T(|\widetilde{H}_j|);\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)-\Psi\left(x-2\varrho_T(|\widetilde{H}_j|);\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) \\
&\quad + P\left[\left|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\right| > 2\varrho_T(|\widetilde{H}_j|) \mid \mathbf{Z},\{\tilde{Z}_t\}_{t=1}^T\right] \\
&\overset{\text{(ii)}}{\leq} 4\xi_T\varrho_T(m+2k) \\
&\quad + P\left[\left|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\right| > 2\varrho_T(|\widetilde{H}_j|) \mid \mathbf{Z},\{\tilde{Z}_t\}_{t=1}^T\right],
\end{aligned}
$$

where (i) follows by the elementary inequality $|\mathbf{1}\{x\leq z\}-\mathbf{1}\{y\leq z\}|\leq \mathbf{1}\{|y-z|\leq|x-y|\}$ for any $x,y,z\in\mathbb{R}$ and (ii) follows by $|\widetilde{H}_j|\leq m+2k$ and the definition of $\mathcal{Q}_j$. Since the above bound does not depend on x, it holds uniformly in $x\in\mathbb{R}$ on the event $\mathcal{Q}_j$. Since $\Psi(\cdot,\cdot)$ is also bounded by one, we have that

$$
\begin{aligned}
&E\left(\sup_{x\in\mathbb{R}}\left|\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)-\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right|\right) \\
&\leq 4\xi_T\varrho_T(m+2k) \\
&\qquad + 2P(\mathcal{Q}_j^c)+P\left[\left|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\right| > 2\varrho_T(|\widetilde{H}_j|)\right] \\
&\leq 4\xi_T\varrho_T(m+2k)+2P(\mathcal{Q}_j^c) \\
&\qquad + P\left[\left|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z})\right)\right| > \varrho_T(|\widetilde{H}_j|)\right] \\
&\qquad + P\left[\left|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)-\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z})\right)\right| > \varrho_T(|\widetilde{H}_j|)\right] \\
&\overset{\text{(i)}}{\leq} 4\xi_T\varrho_T(m+2k)+2P(\mathcal{Q}_j^c)+2\gamma_{1,T} \overset{\text{(ii)}}{\leq} 4\xi_T\varrho_T(m+2k)+2\gamma_{1,T}+2\gamma_{2,T},
\end{aligned}
$$

where (i) follows by Assumption 4 and the fact that $|\widetilde{H}_j|=|\widetilde{H}_R|$ and (ii) follows by $P(\mathcal{Q}_j)\geq 1-\gamma_{2,T}$. Since the above bound holds for all $1\leq j\leq R$, we have

$$
E\left(R^{-1}\sum_{j=1}^{R}\sup_{x\in\mathbb{R}}\left|\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)-\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right|\right)\leq 4\xi_T\varrho_T(m+2k)+2\gamma_{1,T}+2\gamma_{2,T}. \tag{33}
$$

Step 3: bound $\sup_{x\in\mathbb{R}}\left|\hat{F}(x)-\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right|$.

By (30), we notice that

$$\sup_{x\in\mathbb{R}}\left|T\hat{F}(x)-m\sum_{j=1}^{R}\tilde{F}_j(x)\right|=\sup_{x\in\mathbb{R}}\Bigg|\sum_{t=T_0-mR+1}^{T_0}\mathbf{1}\left\{\phi\left(Z_t,...,Z_{t+T_*-1};\hat{\beta}(\mathbf{Z})\right)\leq x\right\}$$
$$+\sum_{t=T_0+1}^{T_0+T_*}\mathbf{1}\left\{\phi\left(Z_{q(t)},...,Z_{q(t+T_*-1)};\hat{\beta}(\mathbf{Z})\right)\leq x\right\}\Bigg|\leq T_*+(T_0-mR)\leq T_*+R-1.$$

Moreover, by (32) and (33), we have that

$$E\left(\sup_{x\in\mathbb{R}}\left|R^{-1}\sum_{j=1}^{R}\tilde{F}_j(x)-\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right)\right|$$
$$\leq C_2\left(\xi_T\varrho_T(m+2k)+m^{1/2}\beta_{\text{mixing}}(m_1-T_*+1)+\sqrt{m_1/m}+\beta_{\text{mixing}}(k-T_*+1)+\gamma_{1,T}+\gamma_{2,T}\right)$$

for some universal constant $C_2>0$.

The above two displays imply that

$$\sup_{x\in\mathbb{R}}\left|\frac{T}{mR}\hat{F}(x)-\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right|\leq\frac{T_*+R-1}{mR}$$
$$+C_2\left(\xi_T\varrho_T(m+2k)+m^{1/2}\beta_{\text{mixing}}(m_1-T_*+1)+\sqrt{m_1/m}+\beta_{\text{mixing}}(k-T_*+1)+\gamma_{1,T}+\gamma_{2,T}\right).$$

Since $\hat{F}(x)\in[0,1]$, we have that

$$\sup_{x\in\mathbb{R}}|(1-T/(mR))\hat{F}(x)|\leq\frac{T}{mR}-1\leq\frac{T-mR}{mR}\leq\frac{T_*+R-1}{mR}.$$

Since $mR\geq T_0/2$ (due to $R<T_0/2$), we have $(T_*+R-1)/(mR)\leq 2T_*T_0^{-1}+m^{-1}\lesssim\sqrt{m_1/m}$. Hence, the above two displays imply that for some universal constant $C_3>0$,

$$E\left(\sup_{x\in\mathbb{R}}\left|\hat{F}(x)-\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right|\right)$$
$$\leq C_3\left(\xi_T\varrho_T(m+2k)+m^{1/2}\beta_{\text{mixing}}(m_1-T_*+1)+\sqrt{m_1/m}+\beta_{\text{mixing}}(k-T_*+1)+\gamma_{1,T}+\gamma_{2,T}\right).\tag{34}$$

Step 4: derive the desired result.

Let A_R be defined as in Step 1 with $j=R$. Following Step 1, we can construct random elements $\{\bar{Z}_t\}_{t\in A_R}$ (on an enlarged probability space) such that (1) $\{\bar{Z}_t\}_{t\in A_R}$ is independent of $\mathbf{Z}_{\widetilde{H}_R}$, (2) $\{\bar{Z}_t\}_{t\in A_R}\overset{d}{=}\{Z_t\}_{t\in A_R}$ and (3) $P(\{\bar{Z}_t\}_{t\in A_R}\neq\{Z_t\}_{t\in A_R})\leq\beta_{\text{mixing}}(k-T_*+1)$.

Define $\bar{G}(\mathbf{Z}_{\widetilde{H}_R})=\phi\left(\bar{Z}_{T_0+1},...,\bar{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)$. Since $\{T_0+1,...,T_0+T_*\}\subset A_R$, we have that $(\bar{Z}_{T_0+1}, ..., \bar{Z}_{T_0+T_*})$ is independent of $\mathbf{Z}_{\widetilde{H}_R}$, which means that

$$P\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}) \leq x \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) = \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \qquad \forall x \in \mathbb{R}.$$

Therefore,

$$\text{conditional on } \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R}),\ \Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \text{ has uniform distribution on } (0,1). \tag{35}$$

We also introduce the following notations to simplify the argument: $\bar{G}(\mathbf{Z}) = \phi\left(\bar{Z}_{T_0+1}, ..., \bar{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z})\right)$ and $G(\mathbf{Z}) = \phi\left(Z_{T_0+1}, ..., Z_{T_0+T_*}; \hat{\beta}(\mathbf{Z})\right)$.

For arbitrary $\alpha \in (0,1)$ and $c > 0$, we observe that

$$\begin{aligned}
&\left|P\left(\hat{F}\left(G(\mathbf{Z})\right) < \alpha \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha\right| \\
&\overset{\text{(i)}}{=} \left|E\left(\mathbf{1}\left\{\hat{F}(G(\mathbf{Z})) < \alpha\right\} \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - E\left(\mathbf{1}\left\{\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) < \alpha\right\} \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right| \\
&\leq E\left(\left|\mathbf{1}\left\{\hat{F}(G(\mathbf{Z})) < \alpha\right\} - \mathbf{1}\left\{\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) < \alpha\right\}\right| \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \\
&\overset{\text{(ii)}}{\leq} E\left(\mathbf{1}\left\{\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha\right| \leq \left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(G(\mathbf{Z}))\right|\right\} \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \\
&= P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha\right| \leq \left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(G(\mathbf{Z}))\right| \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \\
&\leq P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha\right| \leq c \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \\
&\qquad + P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(G(\mathbf{Z}))\right| > c \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \\
&\overset{\text{(iii)}}{\leq} 2c + P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(G(\mathbf{Z}))\right| > c \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)
\end{aligned}$$

where (i) follows by (35), (ii) follows by the elementary inequality $|\mathbf{1}\{x < z\} - \mathbf{1}\{y < z\}| \leq \mathbf{1}\{|y - z| \leq |x - y|\}$ for any $x, y, z \in \mathbb{R}$ and (iii) follows by (35). Now we take expectation on both sides, obtaining

$$\begin{aligned}
&E\left|P\left(\hat{F}\left(G(\mathbf{Z})\right) < \alpha \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha\right| \\
&\leq 2c + P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(G(\mathbf{Z}))\right| > c\right) \\
&\leq 2c + P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(\bar{G}(\mathbf{Z}))\right| > c\right) + P\left(\{\bar{Z}_t\}_{t \in A_R} \neq \{Z_t\}_{t \in A_R}\right) \\
&\leq 2c + P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(\bar{G}(\mathbf{Z}))\right| > c\right) + \beta_{\text{mixing}}(k - T_* + 1) \\
&\leq 2c + c^{-1} E\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(\bar{G}(\mathbf{Z}))\right| + \beta_{\text{mixing}}(k - T_* + 1)
\end{aligned} \tag{36}$$

Define the event

$$\mathcal{A} = \left\{\sup_{x \in \mathbb{R}} \left|\partial\Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) / \partial x\right| \leq \xi_T\right\} \bigcap \left\{\left|\bar{G}(\mathbf{Z}) - \bar{G}(\mathbf{Z}_{\widetilde{H}_R})\right| \leq \varrho_T(|\widetilde{H}_R|)\right\}.$$

By Assumptions 4 and 5, $P(\mathcal{A}^c) \le \gamma_{1,T} + \gamma_{2,T}$. Therefore,

$$
\begin{aligned}
& E\left|\Psi\left(\bar{G}(\mathbf{Z}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right| \\
& = E\left(\left|\Psi\left(\bar{G}(\mathbf{Z}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right| \times \mathbf{1}_{\mathcal{A}}\right) \\
& \qquad\qquad + E\left(\left|\Psi\left(\bar{G}(\mathbf{Z}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right| \times \mathbf{1}_{\mathcal{A}^c}\right) \\
& \overset{\text{(i)}}{\le} E\left(\left|\Psi\left(\bar{G}(\mathbf{Z}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right| \times \mathbf{1}_{\mathcal{A}}\right) + 2P\left(\mathcal{A}^c\right) \\
& \le \xi_T \varrho_T(|\widetilde{H}_R|) + 2P\left(\mathcal{A}^c\right) \le \xi_T \varrho_T(m+2k) + 2\gamma_{1,T} + 2\gamma_{2,T},
\end{aligned}
$$

where (i) follows by the fact that $\Psi(\cdot,\cdot) \in [0,1]$. Hence, we have that

$$
\begin{aligned}
& E\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(\bar{G}(\mathbf{Z}))\right| \\
& \le E\left|\Psi\left(\bar{G}(\mathbf{Z}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(\bar{G}(\mathbf{Z}))\right| + E\left|\Psi\left(\bar{G}(\mathbf{Z}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right| \\
& \le E \sup_{x\in\mathbb{R}}\left|\Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(x)\right| + \xi_T \varrho_T(m+2k) + 2\gamma_{1,T} + 2\gamma_{2,T} \\
& \overset{\text{(i)}}{\le} C_4\left(\xi_T \varrho_T(m+2k) + m^{1/2}\beta_{\text{mixing}}(m_1 - T_* + 1) + \sqrt{m_1/m} + \beta_{\text{mixing}}(k - T_* + 1) + \gamma_{1,T} + \gamma_{2,T}\right)
\end{aligned}
$$

for a universal constant $C_4 > 0$, where (i) follows by (34).

Now we combine (36) and the above display. We also choose

$$
c \asymp \sqrt{\xi_T \varrho_T(m+2k) + m^{1/2}\beta_{\text{mixing}}(m_1 - T_* + 1) + \sqrt{m_1/m} + \beta_{\text{mixing}}(k - T_* + 1) + \gamma_{1,T} + \gamma_{2,T}}.
$$

Then we can find a universal constant $C_5 > 0$ such that

$$
\begin{aligned}
& E\left|P\left(\hat{F}\left(G(\mathbf{Z})\right) < \alpha \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha\right| \\
& \le C_5\sqrt{\xi_T \varrho_T(m+2k) + m^{1/2}\beta_{\text{mixing}}(m_1 - T_* + 1) + \sqrt{m_1/m} + \beta_{\text{mixing}}(k - T_* + 1) + \gamma_{1,T} + \gamma_{2,T}}.
\end{aligned}
$$

Now we choose m_1 satisfying $m_1 \asymp (\log m)^{1/D_3}$ and $m^{1/2}\beta_{\text{mixing}}(m_1 - T_* + 1) \lesssim m^{-1}$. Hence, for some universal constant $C_6 > 0$,

$$
\begin{aligned}
& E\left|P\left(\hat{F}\left(G(\mathbf{Z})\right) < \alpha \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha\right| \\
& \le C_6\sqrt{\xi_T \varrho_T(m+2k)} + C_6\left(m^{-1}(\log m)^{1/D_3}\right)^{1/4} + C_6\sqrt{\beta_{\text{mixing}}(k - T_* + 1)} + C_6\sqrt{\gamma_{1,T}} + C_6\sqrt{\gamma_{2,T}}.
\end{aligned}
$$

Since $m \asymp T_0/R$, the desired result follows once we notice that $\hat{p} \ge 1-\alpha$ and $\hat{F}\left(G(\mathbf{Z})\right) < \alpha$ are the same event.

A.3.1 Proof of Lemma 16

Define $K = \lfloor T/m \rfloor$ and $\hat{F}(x) = m^{-1/2} \sum_{r=1}^{m} \hat{F}_r(x)$, where $\hat{F}_r(x) = K^{-1/2} \sum_{j=1}^{K} [\mathbf{1}\{W_{(j-1)m+r} \leq x\} - G(x)]$ for $1 \leq r \leq m$. Let $\Delta(x) = \sum_{t=mK+1}^{T} [\mathbf{1}\{W_t \leq x\} - G(x)]$. Let $L_T(x) = T^{-1/2} \sum_{t=1}^{T} [\mathbf{1}\{W_t \leq x\} - G(x)]$. Notice that

$$\sqrt{T} L_T(x) = \sqrt{mK} \hat{F}(x) + \Delta(x).$$

Since $|\mathbf{1}\{W_t \leq x\} - G(x)| \leq 1$, it follows that $\sup_{x \in \mathbb{R}} |\Delta(x)| \leq T - mK \leq m - 1$ and thus

$$\sup_{x \in \mathbb{R}} \left| \sqrt{T} L_T(x) - \sqrt{mK} \hat{F}(x) \right| \leq m - 1. \tag{37}$$

By Berbee's coupling (e.g., Lemma 7.1 of Chen et al. (2016)), we can enlarge the probability space and define random variables $\{\bar{W}_t\}_{t=1}^{mK}$ such that (1) $\bar{W}_t \overset{d}{=} W_t$ for all $1 \leq t \leq mT$, (2) $\bar{W}_{(j-1)m+r}$ is independent across $1 \leq j \leq K$ for all r and (3) $P(\bigcup_{t=1}^{mK} \{\bar{W}_t \neq W_t\}) \leq mK\beta_{\text{mixing}}(m) \leq T\beta_{\text{mixing}}(m)$.

We now define $\bar{F}(x) = m^{-1/2} \sum_{r=1}^{m} \bar{F}_r(x)$, where $\bar{F}_r(x) = K^{-1/2} \sum_{j=1}^{K} [\mathbf{1}\{\bar{W}_{(j-1)m+r} \leq x\} - G(x)]$.

Since $\{\bar{W}_{(j-1)m+r}\}_{j=1}^{K}$ is independent, it follows by Dvoretzky-Kiefer-Wolfowitz inequality that for any $z > 0$,

$$P\left(\sup_{x \in \mathbb{R}} |\bar{F}_r(x)| > z \right) \leq 2 \exp(-2z^2).$$

Therefore, we have that

$$E\left(\sup_{x \in \mathbb{R}} |\bar{F}_r(x)| \right) = \int_0^\infty P\left(\sup_{x \in \mathbb{R}} |\bar{F}_r(x)| > z \right) dz \leq 2 \int_0^\infty \exp(-2z^2) dz = \sqrt{\pi/2}.$$

Hence, we have that

$$E\left(\sup_{x \in \mathbb{R}} |\bar{F}(x)| \right) \leq m^{-1/2} \sum_{r=1}^{m} E\left(\sup_{x \in \mathbb{R}} |\bar{F}_r(x)| \right) \leq \sqrt{\pi m/2}.$$

Since $\bar{F}(\cdot) = \hat{F}(\cdot)$ with probability at least $1 - T\beta_{\text{mixing}}(m)$, we have that

$$E\left(\sup_{x \in \mathbb{R}} |\bar{F}(x) - \hat{F}(x)| \right) \leq 2T\beta_{\text{mixing}}(m).$$

Therefore, $E\left(\sup_{x \in \mathbb{R}} |\hat{F}(x)| \right) \leq 2T\beta_{\text{mixing}}(m) + \sqrt{\pi m/2}$.

By (37) and $mK/T \leq 1$, we have that

$$E\left(\sup_{x \in \mathbb{R}} |L_T(x)| \right) \leq 2T\beta_{\text{mixing}}(m) + \sqrt{\pi m/2} + (m-1)/\sqrt{T}.$$

The proof is complete.

A.4 Proof of Lemma 1

By the i.i.d. or exchangeability property of data, we have that

$$\underbrace{\{g(Z_t, \hat{\beta}(\{Z_t\}_{t=1}^T))\}_{t=1}^T}_{\{\hat{u}_t\}_{t=1}^T} \overset{d}{=} \{g(Z_{\pi(t)}, \hat{\beta}(\{Z_{\pi(t)}\}_{t=1}^T))\}_{t=1}^T.$$

Since $\hat{\beta}(\{Z_{\pi(t)}\}_{t=1}^T)$ does not depend on π, we have

$$\{g(Z_{\pi(t)}, \hat{\beta}(\{Z_{\pi(t)}\}_{t=1}^T))\}_{t=1}^T = \underbrace{\{g(Z_{\pi(t)}, \hat{\beta}(\{Z_t\}_{t=1}^T))\}_{t=1}^T}_{\{\hat{u}_{\pi(t)}\}_{t=1}^T}.$$

Therefore, $\{\hat{u}_{\pi(t)}\}_{t=1}^T \overset{d}{=} \{\hat{u}_t\}_{t=1}^T$.

A.5 Proof of Lemma 2

Let X_{jt} denote the (j,t) entry of the matrix $X \in \mathbb{R}^{T\times J}$. We assume the following conditions hold: (1) $E(u_t X_{jt}) = 0$ for $1 \leq j \leq J$. (2) there exist constants $c_1, c_2 > 0$ such that $E|X_{jt}u_t|^2 \geq c_1$ and $E|X_{jt}u_t|^3 \leq c_2$ for any $1 \leq j \leq J$ and $1 \leq t \leq T$; (3) for each $1 \leq j \leq J$, the sequence $\{X_{jt}u_t\}_{t=1}^T$ is β-mixing and the β-mixing coefficient satisfies that $\beta(t) \leq a_1 \exp(-a_2 t^\tau)$, where $a_1, a_2, \tau > 0$ are constants. (4) there exists a constant $c_3 > 0$ such that $\max_{1\leq j\leq J} \sum_{t=1}^T X_{jt}^2 u_t^2 \leq c_3^2 T$ with probability $1-o(1)$. (5) $\log J = o(T^{4\tau/(3\tau+4)})$ and $w \in \mathcal{W}$. (6) There exists a sequence $\ell_T > 0$ such that $(X_t'\delta)^2 \leq \ell_T \|X\delta\|_2^2/T,$ for all $w+\delta \in \mathcal{W}$ with probability $1-o(1)$ for $T_0+1 \leq t \leq T$ and (7) $\ell_T B_T \to 0$ for $B_T = M[\log(T \vee J)]^{(1+\tau)/(2\tau)} T^{-1/2}$.

Then we claim that under conditions (1)-(5) listed above:

(1) There exist a constant $M > 0$ depending only on K and the constants listed above such that with probability $1-o(1)$

$$\|X(\hat{w}-w)\|_2^2/T \leq B_T = M[\log(T\vee J)]^{(1+\tau)/(2\tau)} T^{-1/2}$$

(2) Moreover, if (6) and (7) also hold, then

$$\frac{1}{T}\sum_{t=1}^T \left(\hat{P}_t^N - P_t^N\right)^2 = o_P(1) \text{ and } \hat{P}_t^N - P_t^N = o_P(1), \text{ for any } T_0+1 \leq t \leq T.$$

The following result is useful in deriving the properties of the ℓ_1-constrained estimator.

Lemma 17. *Suppose that (1) $E(u_tX_{jt}) = 0$ for $1 \leq j \leq J$. (2) $\max_{1\leq j\leq J, 1\leq t\leq T} E|X_{jt}u_t|^3 \leq K_1$ for a constant $K_1 > 0$. (3) $\min_{1\leq j\leq J, 1\leq t\leq T} E|X_{jt}u_t|^2 \geq K_2$ for a constant $K_2 > 0$. (4) For each $1 \leq j \leq J$, $\{X_{jt}u_t\}_{t=1}^T$ is β-mixing and the β-mixing coefficients satisfy $\beta(s) \leq D_1 \exp(-D_2 s^\tau)$ for some constants $D_1, D_2, \tau > 0$. Assume $\log J = o(T^{4\tau/(3\tau+4)})$. Then there exists a constant*

$\kappa > 0$ *depending only on* K_1, K_2, D_1, D_2, τ *such that with probability* $1 - o(1)$

$$\max_{1\le j\le J}\left|\sum_{t=1}^{T} X_{jt}u_t\right| < \kappa[\log(T \vee J)]^{(1+\tau)/(2\tau)} \max_{1\le j\le J}\sqrt{\sum_{t=1}^{T} X_{jt}^2 u_t^2}$$

Proof. Define $W_{j,t} = X_{jt}u_t$. Let $m = \left\lfloor [4D_2^{-1}\log(JT)]^{1/\tau}\right\rfloor$ and $k = \lfloor T/m \rfloor$. For simplicity, we assume for now that T/m is an integer. Define

$$H_i = \{i, m+i, 2m+i, ..., (k-1)m+i\} \qquad \forall 1 \le i \le m.$$

By Berbee's coupling (e.g., Lemma 7.1 of Chen et al. (2016)), there exist a sequence of random variables $\{\tilde{W}_{j,t}\}_{t\in H_i}$ such that (1) $\{\tilde{W}_{j,t}\}_{t\in H_i}$ is independent across t, (2) $\tilde{W}_{j,t}$ has the same distribution as $W_{j,t}$ for $t \in H_i$ and (3) $P\left(\bigcup_{t\in H_i}\{\tilde{W}_{j,t} \neq W_{j,t}\}\right) \le k\beta(m)$.

By assumption, $\max_{j,t} E|X_{jt}u_t|^3$ is uniformly bounded above and $\min_{j,t} E|X_{jt}u_t|^2$ is uniformly bounded away from zero. It follows, by Theorem 7.4 of Peña et al. (2008), that there exist constants $C_0, C_1 > 0$ depending on K_1 and K_2 such that for any $0 \le x \le C_0 k^{2/3}$,

$$P\left(\left|\frac{\sum_{t\in H_i}\tilde{W}_{j,t}}{\sqrt{\sum_{t\in H_i}\tilde{W}_{j,t}^2}}\right| > x\right) \le C_1\left(1 - \Phi(x)\right),$$

where $\Phi(\cdot)$ is the cdf of $N(0,1)$. Therefore, for any $0 \le x \le C_0 k^{2/3}$,

$$\begin{aligned} P\left(\left|\frac{\sum_{t\in H_i} W_{j,t}}{\sqrt{\sum_{t\in H_i} W_{j,t}^2}}\right| > x\right) &\le P\left(\left|\frac{\sum_{t\in H_i}\tilde{W}_{j,t}}{\sqrt{\sum_{t\in H_i}\tilde{W}_{j,t}^2}}\right| > x\right) + P\left(\bigcup_{t\in H_i}\{\tilde{W}_{j,t} \neq W_{j,t}\}\right) \\ &\le C_1\left(1 - \Phi(x)\right) + k\beta(m). \end{aligned} \tag{38}$$

The Cauchy-Schwarz inequality implies

$$\begin{aligned} \left|\sum_{t=1}^{T} W_{j,t}\right| \le \sum_{i=1}^{m}\left|\frac{\sum_{t\in H_i} W_{j,t}}{\sqrt{\sum_{t\in H_i} W_{j,t}^2}}\right|\sqrt{\sum_{t\in H_i} W_{j,t}^2} &\le \sqrt{\sum_{i=1}^{m}\left(\frac{\sum_{t\in H_i} W_{j,t}}{\sqrt{\sum_{t\in H_i} W_{j,t}^2}}\right)^2} \times \sqrt{\sum_{i=1}^{m}\sum_{t\in H_i} W_{j,t}^2} \\ &= \sqrt{\sum_{i=1}^{m}\left(\frac{\sum_{t\in H_i} W_{j,t}}{\sqrt{\sum_{t\in H_i} W_{j,t}^2}}\right)^2} \times \sqrt{\sum_{t=1}^{T} W_{j,t}^2}. \end{aligned}$$

Hence,

$$\left|\frac{\sum_{t=1}^{T} W_{j,t}}{\sqrt{\sum_{t=1}^{T} W_{j,t}^2}}\right| \le \sqrt{\sum_{i=1}^{m}\left(\frac{\sum_{t\in H_i} W_{j,t}}{\sqrt{\sum_{t\in H_i} W_{j,t}^2}}\right)^2}.$$

It follows that for any $0 \leq x \leq C_0 k^{2/3}\sqrt{m}$,

$$
\begin{aligned}
&P\left(\left|\frac{\sum_{t=1}^{T} W_{j,t}}{\sqrt{\sum_{t=1}^{T} W_{j,t}^2}}\right| > x\right) \leq P\left(\sqrt{\sum_{i=1}^{m}\left(\frac{\sum_{t\in H_i} W_{j,t}}{\sqrt{\sum_{t\in H_i} W_{j,t}^2}}\right)^2} > x\right) \\
&= P\left(\sum_{i=1}^{m}\left(\frac{\sum_{t\in H_i} W_{j,t}}{\sqrt{\sum_{t\in H_i} W_{j,t}^2}}\right)^2 > x^2\right) \leq \sum_{i=1}^{m} P\left(\left|\frac{\sum_{t\in H_i} W_{j,t}}{\sqrt{\sum_{t\in H_i} W_{j,t}^2}}\right| > \frac{x}{\sqrt{m}}\right) \\
&\overset{\text{(i)}}{\leq} m\left[C_1\left(1-\Phi(x/\sqrt{m})\right) + k\beta(m)\right] \overset{\text{(ii)}}{\leq} C_1 m\sqrt{\frac{m}{2\pi}}x^{-1}\exp\left(-\frac{x^2}{2m}\right) + D_1 km\exp\left(-D_2 m^{\tau}\right) \\
&< C_1 m^{3/2}x^{-1}\exp\left(-\frac{x^2}{2m}\right) + D_1 T\exp\left(-D_2 m^{\tau}\right)
\end{aligned}
$$

where (i) follows by (38) and (ii) follows by the inequality $1-\Phi(a) \leq a^{-1}\phi(a)$ (with ϕ being the pdf of $N(0,1)$) and $\beta(m) \leq D_1\exp(-D_2 m^{\tau})$.

By the union bound, it follows that for any $0 \leq x \leq C_0 k^{2/3}\sqrt{m}$,

$$
P\left(\max_{1\leq j\leq J}\left|\frac{\sum_{t=1}^{T} W_{j,t}}{\sqrt{\sum_{t=1}^{T} W_{j,t}^2}}\right| > x\right) \leq C_1 J m^{3/2}x^{-1}\exp\left(-\frac{x^2}{2m}\right) + D_1 JT\exp\left(-D_2 m^{\tau}\right).
$$

Now we choose $x = 2\sqrt{m\log(Jm^{3/2})}$. Since $\log J = o(T^{4\tau/(3\tau+4)})$ and $k \asymp T/m$, it can be very easily verified that $x \ll C_0 k^{2/3}\sqrt{m}$ and the two terms on the right-hand side of the above display tend to zero. The desired result follows.

If T/k is not an integer, then we simply add one observation from $\{W_{j,t}\}_{t=km+1}^{T}$ to each of H_i for $1 \leq i \leq m$. The bound in (38) holds with C_1 large enough. The proof is complete. □

Now we are ready to prove Lemma 2.

Proof of Lemma 2. Let $\Delta = \hat{w} - w$. Since $\|w\|_1 \leq K$, we have $\|Y - X\hat{w}\|_2^2 \leq \|Y - Xw\|_2^2$. Notice that $Y - Xw = u$ and $Y - X\hat{w} = u - X\Delta$. Therefore, $\|u - X\Delta\|_2^2 \leq \|u\|_2^2$, which means $\|X\Delta\|_2^2 \leq 2u'X\Delta$. Now we observe that

$$
\|X\Delta\|_2^2 \leq 2u'X\Delta \overset{\text{(i)}}{\leq} 2\|Xu\|_\infty\|\Delta\|_1 \overset{\text{(ii)}}{\leq} 4K\|Xu\|_\infty, \tag{39}
$$

where (i) follows by Hölder's inequality and (ii) follows by $\|\Delta\|_1 \leq 2K$ (since $\|\hat{w}\|_1 \leq K$ and $\|w\|_1 \leq K$). By Lemma 17, there exists a constant $\kappa > 0$ such that

$$
P\left(\max_{1\leq j\leq J}\left|\sum_{t=1}^{T} X_{jt}u_t\right| > \kappa[\log(T\vee J)]^{(1+\tau)/(2\tau)}\max_{1\leq j\leq J}\sqrt{\sum_{t=1}^{T} X_{jt}^2 u_t^2}\right) = o(1).
$$

Since $P\left(\max_{1\leq j\leq J}\sum_{t=1}^{T}X_{jt}^2u_t^2\leq c_3^2T\right)\to 1$, it follows that

$$P\left(\max_{1\leq j\leq J}\left|\sum_{t=1}^{T}X_{jt}u_t\right|>\kappa c_3[\log(T\vee J)]^{(1+\tau)/(2\tau)}\sqrt{T}\right)=o(1). \tag{40}$$

Part (1) follows by combining (39) and (40). Part (2) follows by part (1) and $\ell_T B_T = o(1)$. □

A.6 Proof of Lemma 3

We borrow results and notations from Bai (2003). Following standard notation, we use i instead of j to denote units. Here are the regularity conditions from Bai (2003).

Suppose that there exists a constant $D_0>0$ the following conditions hold:
(1) $\max_{1\leq t\leq T}E\|F_t\|_2^4\leq D_0$, $\max_{1\leq j\leq N}\|\lambda_j\|_2^4\leq D_0$, $\max_{jt}E|u_{jt}|^8\leq D_0$ and $E(u_{jt})=0$.
(2) $\max_s N^{-1}\sum_{t=1}^{T}|\sum_{i=1}^{N}E(u_{is}u_{it})|\leq D_0$ and $\max_i\sum_{j=1}^{N}\max_{1\leq t\leq T}|E(u_{it}u_{jt})|\leq D_0$.
(3) $(NT)^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N}|E(u_{it}u_{js})|\leq D_0$ and $\max_{s,t}E|N^{-1/2}\sum_{i=1}^{N}[u_{is}u_{it}-E(u_{is}u_{it})]|^4\leq D_0$.
(4) $N^{-1}\sum_{i=1}^{N}E\|T^{-1/2}\sum_{t=1}^{T}F_tu_{it}\|_2^2\leq D_0$.
(5) $\max_t E\|(NT)^{-1/2}\sum_{s=1}^{T}\sum_{i=1}^{N}F_s[u_{is}u_{it}-E(u_{is}u_{it})]\|_2^2\leq D_0$.
(6) $E\|(NT)^{-1/2}\sum_{t=1}^{T}\sum_{i=1}^{N}F_t\lambda_i'u_{it}\|_2^2\leq D_0$.

Moreover, we assume the following conditions: (7) for each t, $N^{-1/2}\sum_{i=1}^{N}\lambda_iu_{it}\to^d N(0,\Gamma_t)$ as $N\to\infty$, where $\Gamma_t=\lim_{N\to\infty}N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j'E(u_{it}u_{jt})$. (8) for each i, $T^{-1/2}\sum_{t=1}^{T}F_tu_{it}\to^d N(0,\Phi_i)$ as $T\to\infty$, where $\Phi_i=\lim_{T\to\infty}T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}E(F_tF_s'u_{is}u_{it})$. (9) $N^{-1}\sum_{i=1}^{N}\lambda_i\lambda_i'\to\Sigma_\Lambda$ and $T^{-1}\sum_{t=1}^{T}F_tF_t'=\Sigma_F+o_P(1)$ for some $k\times k$ positive definite matrices Σ_Λ and Σ_F satisfying that $\Sigma_\Lambda\Sigma_F$ has distinct eigenvalues.

What follows below is the proof of the lemma. We recall some notations used by Bai (2003). Let $F=(F_1,\ldots,F_T)'\in\mathbb{R}^{T\times k}$ and $\Lambda=(\lambda_1,\ldots,\lambda_N)'\in\mathbb{R}^{N\times k}$. Define $H=(\Lambda'\Lambda/N)(F'\tilde{F}/T)V_{NT}^{-1}$, where $V_{NT}\in\mathbb{R}^{k\times k}$ is the diagonal matrix with the largest k eigenvalues of $Y^N(Y^N)'/(NT)$ on the diagonal and $\tilde{F}$ is the normalized F, namely $\tilde{F}'\tilde{F}/T=I_k$.

We start with the first equation in the proof of Theorem 3 in Bai (2003) (on page 166):

$$\hat{\lambda}_1'\hat{F}_t-\lambda_1'F_t=\left(\hat{F}_t-H'F_t\right)'H^{-1}\lambda_1+\hat{F}_t'(\hat{\lambda}_1-H^{-1}\lambda_1). \tag{41}$$

The rest of the proof proceeds in two steps. We first recall some results from Bai (2003) and then derive the desired result.

Step 1: recall useful results from Bai (2003). By Lemma A.1 of Bai (2003),

$$\sum_{t=1}^{T}\|\hat{F}_t-H'F_t\|_2^2=O_P(T/\delta_{NT}^2), \tag{42}$$

where $\delta_{NT} = \min\{\sqrt{N}, \sqrt{T}\}$. By definition, $\hat{F}'\hat{F}/T = I_k$, which means

$$\sum_{t=1}^{T} \|\hat{F}_t\|_2^2 = \sum_{t=1}^{T} \mathrm{trace}(\hat{F}_t \hat{F}_t') = \mathrm{trace}(\hat{F}'\hat{F}) = kT. \tag{43}$$

By Theorem 2 of Bai (2003),

$$\hat{\lambda}_1 = H^{-1}\lambda_1 + O_P(\max\{T^{-1/2}, N^{-1}\}). \tag{44}$$

By the proof of part (i) in Theorem 2 of Bai (2003), H converges in probability to a nonsingular matrix; see page 166 of Bai (2003). Hence, $\|H^{-1}\| = O_P(1)$. By assumption, $\|\lambda_1\|_2 = O(1)$. Hence,

$$\|H^{-1}\lambda_1\|_2 = O_P(1). \tag{45}$$

Step 2: prove the desired result.

Therefore,

$$\begin{aligned}
\sum_{t=1}^{T} \left(\hat{\lambda}_1' \hat{F}_t - \lambda_1' F_t\right)^2 &\overset{\text{(i)}}{\leq} 2\sum_{t=1}^{T} \left[\left(\hat{F}_t - H'F_t\right)' H^{-1}\lambda_1\right]^2 + 2\sum_{t=1}^{T} \left[\hat{F}_t'(\hat{\lambda}_1 - H^{-1}\lambda_1)\right]^2 \\
&\leq 2\sum_{t=1}^{T} \|\hat{F}_t - H'F_t\|_2^2 \times \|H^{-1}\lambda_1\|_2^2 + 2\sum_{t=1}^{T} \|\hat{F}_t\|_2^2 \times \|\hat{\lambda}_1 - H^{-1}\lambda_1\|_2^2 \\
&\overset{\text{(ii)}}{=} O_P(T/\delta_{NT}^2) \times O_P(1) + 2kT \times O_P(\max\{T^{-1}, N^{-2}\}) \\
&= O_P(T/\delta_{NT}^2),
\end{aligned}$$

where (i) follows by (41) and the elementary inequality of $(a+b)^2 \leq 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$ and (ii) follows by (42), (43), (44) and (45). Since $n = |\Pi| = T$ for moving block permutation, we have

$$\frac{1}{n}\sum_{t=1}^{T} \left(\hat{\lambda}_1' \hat{F}_t - \lambda_1' F_t\right)^2 = O_P\left(\frac{1}{\min\{N, T\}}\right).$$

This proves part (1) of condition (A).

Notice that Theorem 3 of Bai (2003) implies $\hat{\lambda}_1'\hat{F}_t - \lambda_1' F_t = O_P(1/\delta_{NT})$. Part (2) of Condition (A) follows. The proof is complete.

A.7 Proof of Lemma 4

We recite conditions from Bai (2009). Following standard notation, we use i instead of j to denote units.

Suppose that there exists a constant $D_1 > 0$ the following conditions hold:

(1) $\max_{i,t} E\|X_{it}\|_2^4 \leq D_1$, $\max_t E\|F_t\|_2^4 \leq D_1$, $\max_i E\|\lambda_i\|_2^4 \leq D_1$ and $\max_{i,t} E|u_{it}|^8 \leq D_1$.

(2) $N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\max_{t,s}|E(u_{it}u_{js})| \leq D_1$and $T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}\max_{i,j}|E(u_{it}u_{js})| \leq D_1$.

(3) $(NT)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{s=1}^{T}\sum_{t=1}^{T}|E(u_{it}u_{js})| \leq D_1$.

(4) $\max_{t,s} E\left|N^{-1/2}\sum_{i=1}^{N}[u_{is}u_{it}-E(u_{is}u_{it})]\right|^4 \leq D_1$.
(5) $T^{-2}N^{-1}\sum_{t,s,q,v}\sum_{i,j}|cov(u_{it}u_{ts},u_{jq}u_{jv})| \leq D_1$
(6) $T^{-1}N^{-2}\sum_{t,s}\sum_{i,j,k,q}|cov(u_{it}u_{jt},u_{ks}u_{qs})| \leq D_1$.
(7) the largest eigenvalue of $E(u_i u_i')$ is bounded by D_1, where $u_i = (u_{i1},...,u_{iT})' \in \mathbb{R}^T$.

Moreover, the following conditions also hold: (8) $u = (u_1,\ldots,u_N)$ is independent of (X,F,Λ). (9) $F'F/T = \Sigma_F + o_P(1)$ and $\Lambda'\Lambda/N = \Sigma_\Lambda + o_P(1)$ for some matrices Σ_F and Σ_Λ. (10) N/T is bounded away from zero and infinity. (11) For $X_i = (X_{i1},...,X_{iT})' \in \mathbb{R}^{T\times k_x}$ and $M_F = I_T - F(F'F)^{-1}F'$, we have

$$\inf_{F:\ F'F/T=I_k}\frac{1}{NT}\sum_{i=1}^{N}X_i'M_FX_i - \frac{1}{T}\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}X_i'M_FX_j\lambda_i'(\Lambda'\Lambda/N)^{-1}\lambda_j\right] > 0.$$

What follows below is the proof of the lemma. We introduce some notations used in Bai (2009). Let $H = (\Lambda'\Lambda/N)(F'\hat{F}/T)V_{NT}^{-1}$, where V_{NT} is the diagonal matrix that contains the k largest eigenvalues of $(NT)^{-1}\sum_{i=1}^{N}(Y_i^N - X_i\hat{\beta})(Y_i^N - X_i\hat{\beta})'$ with $Y_i^N = (Y_{i1}^N, Y_{i2}^N, ..., Y_{iT}^N)' \in \mathbb{R}^T$. Let $\delta_{NT} = \min\{\sqrt{N},\sqrt{T}\}$. The rest of the proof proceeds in two steps. We first derive bounds for $\sum_{t=1}^{T}(\hat{u}_{1t}-u_{1t})^2$ and then prove the pointwise result.

Step 1: derive bounds for $\sum_{t=1}^{T}(\hat{u}_{1t}-u_{1t})^2$.

Define $\Delta_\beta = \hat{\beta}-\beta$ and $\Delta_{F,t} = \hat{F}_t - H'F_t$. Denote $\Delta_F = (\Delta_{F,1},...,\Delta_{F,T})' \in \mathbb{R}^{T\times k}$. Notice that $\hat{F} - FH = \Delta_F$. As pointed out on page 1237 of Bai (2009),

$$\hat{\lambda}_1 = T^{-1}\hat{F}'(Y_1^N - X_1\hat{\beta}) = T^{-1}\hat{F}'(u_1 + F\lambda_1 - X_1\Delta_\beta). \tag{46}$$

Notice that

$$\begin{aligned}
&|\hat{u}_{1t}-u_{1t}|^2 = \left|F_t'\lambda_1 - \hat{F}_t'\hat{\lambda}_1 - X_{1t}'\Delta_\beta\right|^2\\
&\overset{\text{(i)}}{=} \left|F_t'\lambda_1 - T^{-1}(H'F_t+\Delta_{F,t})'\hat{F}'(u_1+F\lambda_1-X_1\Delta_\beta) - X_{1,t}'\Delta_\beta\right|^2\\
&\leq \left[\left|F_t'\left(I_k - H\hat{F}'F/T\right)\lambda_1\right| + \left|T^{-1}\Delta_{F,t}'\hat{F}'F\lambda_1\right| + \left|T^{-1}\hat{F}_t'\hat{F}'(u_1-X_1\Delta_\beta)\right| + |X_{1t}'\Delta_\beta|\right]^2\\
&\lesssim \left[F_t'\left(I_k - H\hat{F}'F/T\right)\lambda_1\right]^2 + \left[T^{-1}\Delta_{F,t}'\hat{F}'F\lambda_1\right]^2 + \left[T^{-1}\hat{F}_t'\hat{F}'(u_1-X_1\Delta_\beta)\right]^2 + [X_{1t}'\Delta_\beta]^2,
\end{aligned} \tag{47}$$

where (i) follows by (46) and $\hat{F}_t = H'F_t + \Delta_{F,t}$. Therefore,

$$\begin{aligned}
\sum_{t=1}^{T}(\hat{u}_{1t}-u_{1t})^2 &\lesssim \sum_{t=1}^{T}\left[F_t'\left(I_k - H\hat{F}'F/T\right)\lambda_1\right]^2 + \sum_{t=1}^{T}\left[T^{-1}\Delta_{F,t}'\hat{F}'F\lambda_1\right]^2\\
&\quad + \sum_{t=1}^{T}\left[T^{-1}\hat{F}_t'\hat{F}'(u_1-X_1\Delta_\beta)\right]^2 + \sum_{t=1}^{T}[X_{1t}'\Delta_\beta]^2\\
&\overset{\text{(i)}}{=} \lambda_1'\left(I_k - H\hat{F}'F/T\right)'(F'F)\left(I_k - H\hat{F}'F/T\right)\lambda_1
\end{aligned}$$

$$
\begin{aligned}
&+ T^{-2}\left(\hat{F}'F\lambda_1\right)'(\Delta_F'\Delta_F)\left(\hat{F}'F\lambda_1\right) + T^{-1}\left\|\hat{F}'(u_1 - X_1\Delta_\beta)\right\|_2^2 + \|X_1\Delta_\beta\|_2^2 \\
&\overset{\text{(ii)}}{=} O_P\left(T\|\Delta_\beta\|_2^2 + T\delta_{NT}^{-4}\right) + O_P\left(T\|\Delta_\beta\|_2^2 + T\delta_{NT}^{-2}\right) + O_P\left(1 + T\delta_{NT}^{-4} + T\|\Delta_\beta\|_2^2\right) + O_P(T\|\Delta_\beta\|_2^2) \\
&= O_P\left(1 + T\|\Delta_\beta\|_2^2 + T\delta_{NT}^{-2}\right),
\end{aligned}
$$

where (i) follows by $\sum_{t=1}^T \hat{F}_t\hat{F}_t' = \hat{F}'\hat{F} = TI_k$ and (ii) follows by Lemma 18, together with $\|F\| = O_P(\sqrt{T})$, $\lambda_1 = O(1)$ and $\|\hat{F}\| = O_P(\sqrt{T})$. Since $N \asymp T$, Theorem 1 of Bai (2009) implies $\|\Delta_\beta\|_2 = O_P(1/\sqrt{NT}) = O_P(T^{-1})$. Therefore, the above display implies

$$
\sum_{t=1}^T (\hat{u}_{1t} - u_{1t})^2 = O_P(1).
$$

Step 2: show the pointwise result.
By (47), we have

$$
\begin{aligned}
|\hat{u}_{1t} - u_{1t}| &\leq \left|F_t'\left(I_k - H\hat{F}'F/T\right)\lambda_1\right| + \left|T^{-1}\Delta_{F,t}'\hat{F}'F\lambda_1\right| + \left|T^{-1}\hat{F}_t'\hat{F}'(u_1 - X_1\Delta_\beta)\right| + |X_{1t}'\Delta_\beta| \\
&\overset{\text{(i)}}{\leq} \|F_t\|_2 \cdot \|\lambda_1\|_2 \cdot O_P\left(\|\Delta_\beta\|_2 + \delta_{NT}^{-2}\right) + O_P\left(T\|\Delta_\beta\|_2 + T\delta_{NT}^{-2}\right) \cdot T^{-1}\|F\lambda_1\|_2 \\
&\quad + T^{-1}\|\hat{F}_t\|_2 \cdot O_P\left(\sqrt{T} + T\delta_{NT}^{-2} + T\|\Delta_\beta\|_2\right) + \|X_{1t}\|_2 \cdot \|\Delta_\beta\|_2 \overset{\text{(ii)}}{\leq} O_P(T^{-1/2}),
\end{aligned}
$$

where (i) follows by $I_k - H\hat{F}'F/T = O_P(\|\Delta_\beta\|_2 + \delta_{NT}^{-2})$, $\|\hat{F}\Delta_{F,t}\| = O_P(T\|\Delta_\beta\|_2 + T\delta_{NT}^{-2})$, and $\|\hat{F}'(u_1 - X_1\Delta_\beta)\|_2 = O_P(\sqrt{T} + T\delta_{NT}^{-2} + T\|\Delta_\beta\|_2)$ (due to Lemma 18), whereas (ii) follows by $\|\hat{F}_t\|_2 = O_P(1)$ (Lemma 18), $\|X_{1t}\|_2 = O_P(1)$, $\|F_t\|_2 = O_P(1)$, $\lambda_1 = O(1)$, $\|\Delta_\beta\|_2 = O_P(T^{-1})$ and $\|F\lambda_1\|_2 = O_P(\sqrt{T})$.

Lemma 18. *Suppose that the assumption of Lemma 4 holds. Let δ_{NT}, H, Δ_F and u_1 be defined as in the proof of Lemma 4. Then (1) $I_k - H\hat{F}'F/T = O_P(\|\Delta_\beta\|_2 + \delta_{NT}^{-2})$; (2) $\Delta_F'\Delta_F = O_P(T\|\Delta_\beta\|_2^2 + T\delta_{NT}^{-2})$; (3) $\left\|\hat{F}'(u_1 - X_1\Delta_\beta)\right\| = O_P\left(\sqrt{T} + T\delta_{NT}^{-2} + T\|\Delta_\beta\|_2\right)$; (4) $\|X_1\Delta_\beta\|_2 = O_P(\sqrt{T}\|\Delta_\beta\|_2)$; (5) $\|\hat{F}\Delta_{F,t}\|_2 = O_P(T\|\Delta_\beta\|_2 + T\delta_{NT}^{-2})$; (6) $\|\hat{F}_t\|_2 = O_P(1)$ for $1 \leq t \leq T$.*

Proof. **Proof of part (1).** Lemma A.7(i) in Bai (2009) implies HH' converges in probability to a nonsingular matrix. Hence,

$$
H = O_P(1) \quad \text{and} \quad H^{-1} = O_P(1). \tag{48}
$$

Notice that

$$
\begin{aligned}
I_k - H\hat{F}'F/T &\overset{\text{(i)}}{=} I_k - H(FH + \Delta_F)'F/T = I_k - (HH')(F'F/T) - H\Delta_F'F/T \\
&\overset{\text{(ii)}}{=} O_P(\|\Delta_\beta\|_2) + O_P(\delta_{NT}^{-2}) - H\Delta_F'F/T \\
&\overset{\text{(iii)}}{=} O_P(\|\Delta_\beta\|_2) + O_P(\delta_{NT}^{-2}),
\end{aligned} \tag{49}
$$

where (i) holds by $\hat{F} = FH + \Delta_F$, (ii) holds by $I_k - (HH')(F'F/T) = O_P(\|\Delta_\beta\|_2) + O_P(\delta_{NT}^{-2})$

(due to Lemma A.7(i) in Bai (2009)) and (iii) holds by (48) and $\Delta_F' F/T = O_P(\|\Delta_\beta\|_2) + O_P(\delta_{NT}^{-2})$ (due to Lemma A.3(i) in Bai (2009)). This proves part (1).

Proof of part (2). Part (2) follows by Proposition A.1 of Bai (2009):

$$T^{-1}\Delta_F'\Delta_F = O_P(\|\Delta_\beta\|_2^2) + O_P(\delta_{NT}^{-2}). \tag{50}$$

Proof of part (3). To see part (3), first observe that the independence between u_1 and F implies that

$$E(\|F'u_1\|^2 \mid F) \leq \sum_{t=1}^{T} E(F_t'F_t u_{1t}^2 \mid F) = \sum_{t=1}^{T} F_t'F_t E(u_{1t}^2).$$

It follows that

$$E\left(\|F'u_1\|^2\right) \leq \sum_{t=1}^{T} E(F_t'F_t)E(u_{1t}^2) \overset{\text{(i)}}{\lesssim} T\sum_{t=1}^{T} E(u_{1t}^2) = O(T),$$

where (i) holds by the uniform boundedness of $E(F_t'F_t)$. This means that

$$\|F'u_1\|_2 = O_P(\sqrt{T}). \tag{51}$$

Notice that

$$\begin{aligned}\left\|\hat{F}'(u_1 - X_1\Delta_\beta)\right\|_2 &\leq \|H'F'u_1\|_2 + \left\|\left(\hat{F} - FH\right)' u_1\right\| + \|\hat{F}\| \cdot \|X_1\| \cdot \|\Delta_\beta\|_2 \\ &\overset{\text{(i)}}{=} \|H'F'u_1\|_2 + \left(O_P(T^{1/2}\|\Delta_\beta\|_2) + O_P(T\delta_{NT}^{-2})\right) + O_P(T\|\Delta_\beta\|_2) \\ &\overset{\text{(ii)}}{=} O_P(\sqrt{T}) + \left(O_P(T^{1/2}\|\Delta_\beta\|_2) + O_P(T\delta_{NT}^{-2})\right) + O_P(T\|\Delta_\beta\|_2),\end{aligned}$$

where (i) follows by $\left(\hat{F} - FH\right)' u_1/T = O_P(T^{-1/2}\|\Delta_\beta\|_2) + O_P(\delta_{NT}^{-2})$ (due to Lemma A.4 in Bai (2009)) and the fact that $\|\hat{F}\| = O(\sqrt{T})$ and $\|X_1\| = O_P(\sqrt{T})$ (see the beginning of Appendix A in Bai (2009)), whereas (ii) follows by (48) and (51). We have proved part (3).

Proof of part (4). We notice that $\|X_1\| = O_P(\sqrt{T})$; see the beginning of Appendix A in Bai (2009). Part (4) follows by $\|X_1\Delta_\beta\| \leq \|X_1\| \cdot \|\Delta_\beta\|_2$.

Proof of part (5). Notice that

$$\|\hat{F}\Delta_{F,t}\|_2/T \leq \|\hat{F}\Delta_F\|/T \overset{\text{(i)}}{\leq} O_P(\|\Delta_\beta\|_2) + O_P(\delta_{NT}^{-2}),$$

where (i) follows by Lemma A.3(ii) in Bai (2009). We have proved part (5).

Proof of part (6). Notice that

$$T^{-1}\|\Delta_{F,t}\|_2^2 \leq T^{-1}\Delta_F'\Delta_F = T^{-1}\hat{F}'\Delta_F - T^{-1}H'F'\Delta_F \overset{\text{(i)}}{=} O_P(\|\Delta_\beta\|_2) + O_P(\delta_{NT}^{-2}),$$

where (i) follows by Lemma A.3(i)-(ii) of Bai (2009). By Theorem 1 of Bai (2009) and by the assumption of $N \asymp T$, we have that $\|\Delta_{F,t}\|_2 = O_P(1)$. By $\|\hat{F}_t\|_2 \leq \|H'F_t\|_2 + \|\Delta_{F,t}\|_2$, $F_t = O_P(1)$ and $H = O_P(1)$, we can see that $\|\hat{F}_t\|_2 = O_P(1)$. The proof is complete. □

A.8 Proof of Lemma 5

We start with the assumptions. Recall $N = J + 1$. Assume that (1) $\{u_j\}_{j=1}^N$ is independent across j conditional on M, (2) $\max_{1 \leq j \leq N} T^{-1} \sum_{t=1}^T E(|u_{jt}|^{2\kappa_1} \mid M) = O_P(1)$ for some constant $\kappa_1 > 1$, (3) $\|N^{-1} \sum_{j=1}^N E(u_j u_j' \mid M)\| = O_P(1)$ and (4) there exists a sequence $\ell_T > 0$ such that
$\ell_T (NT)^{-1} K \sqrt{N \vee (N^{1/\kappa_1} T \log N)} = o(1)$ and with probability $1 - o(1)$, $T^{-1} \sum_{t=1}^T (\hat{M}_{1t} - M_{1t})^2 \leq \ell_T (NT)^{-1} \sum_{t=1}^T \sum_{j=1}^N (\hat{M}_{jt} - M_{jt})^2$ and $(\hat{M}_{1t} - M_{1t})^2 \leq \ell_T (NT)^{-1} \sum_{t=1}^T \sum_{j=1}^N (\hat{M}_{jt} - M_{jt})^2$ for $T_0 + 1 \leq t \leq T$.

We now prove Lemma 5. Define $\Delta = \hat{M} - M$. Let $Y \in \mathbb{R}^{T \times N}$ be the matrix whose (t, j) entry is Y_{jt}^N. For (j, t), define the matrix $Q_{jt} \in \mathbb{R}^{N \times T}$ by $Q_{is} = \mathbf{1}\{(i, s) = (j, t)\}$, i.e, a matrix of zeros except that the (j, t) entry is one. Then we can write the model as

$$Y_{jt}^N = \operatorname{trace}(Q_{jt}' M) + u_{jt} \qquad \text{for} \quad 1 \leq j \leq N, \ 1 \leq t \leq T. \tag{52}$$

Notice that the estimator $\hat{M}$ satisfies

$$\|\hat{M}\|_* \leq K$$

and

$$\sum_{t=1}^T \sum_{j=1}^N \left(Y_{jt}^N - \operatorname{trace}(Q_{jt}' \hat{M}) \right)^2 \leq \sum_{t=1}^T \sum_{j=1}^N \left(Y_{jt}^N - \operatorname{trace}(Q_{jt}' M) \right)^2.$$

Plugging (52) into the above inequality and rearranging terms, we obtain

$$\begin{aligned} \sum_{t=1}^T \sum_{j=1}^N \left(\operatorname{trace}(Q_{jt}' \Delta) \right)^2 \leq 2 \sum_{t=1}^T \sum_{j=1}^N u_{jt} \operatorname{trace}(Q_{jt}' \Delta) &= 2\operatorname{trace}\left(\left[\sum_{t=1}^T \sum_{j=1}^N u_{jt} Q_{jt} \right]' \Delta \right) \\ &\overset{\text{(i)}}{=} 2\operatorname{trace}(u' \Delta) \\ &\overset{\text{(ii)}}{\leq} 2\|u\| \cdot \|\Delta\|_* \\ &\overset{\text{(iii)}}{\leq} 4K\|u\|, \end{aligned} \tag{53}$$

where (i) follows by $\sum_{t=1}^T \sum_{j=1}^N u_{jt} Q_{jt} = u$, (ii) follows by the trace duality property (see e.g., McCarthy (1967), Rotfeld (1969) and Rohde and Tsybakov (2011)) and (iii) follows by the fact that $\|\hat{M}\|_* \leq K$ and $\|M\|_* \leq K$.

To bound $\|u\|$, we apply Lemma 19. Note that the conditions of Lemma 19 are satisfied by our assumption. Therefore, $E(\|u\| \mid M) = O_P\left(\sqrt{N \vee (N^{1/\kappa_1} T \log N)} \right)$. This and (53) imply that

$$(NT)^{-1}\sum_{t=1}^{T}\sum_{j=1}^{N}\left(\mathrm{trace}(Q'_{jt}\Delta)\right)^2 = O_P\left((NT)^{-1}K\sqrt{N\vee(N^{1/\kappa_1}T\log N)}\right).$$

The desired result follows by Assumption (4) listed at the beginning of the proof.

Lemma 19. *Suppose that the following conditions hold:*
(i) $\{u_j\}_{j=1}^N$ *is independent across* j *conditional on* M.
(ii) $\max_{1\le j\le N} T^{-1}\sum_{t=1}^T E(|u_{jt}|^{2\kappa_1}\mid M) = O_P(1)$ *for some constant* $\kappa_1 > 1$.
(iii) $\|N^{-1}\sum_{j=1}^N E(u_j u_j' \mid M)\| = O_P(1)$.

Then $E(\|u\|\mid M) = O_P\left(\sqrt{N\vee(N^{1/\kappa_1}T\log N)}\right)$.

Proof. Recall the elementary inequality $|T^{-1}\sum_{t=1}^T a_t| \le [T^{-1}\sum_{t=1}^T |a_t|^\kappa]^{1/\kappa}$ for any $\kappa > 1$ (due to Liapunov's inequality). It follows that $T^{-1}\sum_{t=1}^T u_{jt}^2 \le [T^{-1}\sum_{t=1}^T |u_{jt}|^{2\kappa_1}]^{1/\kappa_1}$, which means

$$\left(\sum_{t=1}^T u_{jt}^2\right)^{\kappa_1} \le T^{\kappa_1-1}\sum_{t=1}^T |u_{jt}|^{2\kappa_1}. \tag{54}$$

Hence,

$$\begin{aligned}
E\left(\left[\max_{1\le j\le N}\sum_{t=1}^T u_{jt}^2\right]\mid M\right) &\overset{\text{(i)}}{\le} \left\{E\left[\max_{1\le j\le N}\left(\sum_{t=1}^T u_{jt}^2\right)^{\kappa_1}\mid M\right]\right\}^{1/\kappa_1}\\
&\le \left\{E\left[\sum_{i=1}^N\left(\sum_{t=1}^T u_{jt}^2\right)^{\kappa_1}\mid M\right]\right\}^{1/\kappa_1}\\
&\overset{\text{(ii)}}{\le} \left\{E\left[T^{\kappa_1-1}\sum_{j=1}^N\sum_{t=1}^T |u_{jt}|^{2\kappa_1}\mid M\right]\right\}^{1/\kappa_1}\\
&\le \left\{\left[NT^{\kappa_1}\max_{1\le i\le N} T^{-1}\sum_{t=1}^T E\left(|u_{jt}|^{2\kappa_1}\mid M\right)\right]\right\}^{1/\kappa_1}\\
&\overset{\text{(iii)}}{=} \{[NT^{\kappa_1}O_P(1)]\}^{1/\kappa_1} = O_P(N^{1/\kappa_1}T),
\end{aligned}$$

where (i) follows by Liapunov's inequality, (ii) follows by (54) and (iii) follows by the assumption that $\max_{1\le j\le N} T^{-1}\sum_{t=1}^T E(|u_{jt}|^{2\kappa_1}\mid M) = O_P(1)$. Therefore, it follows, by Theorem 5.48 and Remark 5.49 of Vershynin (2010), that

$$\begin{aligned}
E\left(\|u\|\mid M\right) \le \sqrt{E\left(\|u\|^2\mid M\right)} &\le \|E(u'u\mid M)/N\|^{1/2}\sqrt{N}\\
&\quad + O\left(\sqrt{O(N^{1/\kappa_1}T)\log\min\left(O(N^{1/\kappa_1}T),N\right)}\right)\\
&\overset{\text{(i)}}{\le} O_P\left(\sqrt{N}\right) + O\left(\sqrt{N^{1/\kappa_1}T\log N}\right),
\end{aligned}$$

where (i) holds by the assumption of $\|E(u'u \mid M)/N\| = \|N^{-1}\sum_{j=1}^{N} E(u_j u_j' \mid M)\| = O_P(1)$. The proof is complete. □

A.9 Proof of Lemma 6

By the analysis on page 215-216 of Hamilton (1994) (leading to Equation (8.2.29) therein), we have that $\hat{\rho} - \rho = o_P(1)$. Hence,

$$\sum_{t=K+1}^{T} (\hat{u}_t - u_t)^2 = \sum_{t=K+1}^{T} (y_t'(\rho - \hat{\rho}))^2 = (\hat{\rho} - \rho)' \left(\sum_{t=K+1}^{T} y_t y_t' \right) (\hat{\rho} - \rho) \leq \|\hat{\rho} - \rho\|_2^2 \times \left\| \sum_{t=K+1}^{T} y_t y_t' \right\|.$$

The analysis on page 215 of Hamilton (1994) (leading to Equation (8.2.26) therein) implies that

$$T^{-1} \sum_{t=K+1}^{T} y_t y_t' = E(y_t y_t') + o_P(1),$$

which means $\left\| \sum_{t=K+1}^{T} y_t y_t' \right\| = O_P(T)$. Since $\hat{\rho} - \rho = o_P(1)$, the above display implies that

$$\sum_{t=K+1}^{T} (\hat{u}_t - u_t)^2 = o_P(T).$$

Since $\hat{u}_t - u_t = y_t'(\rho - \hat{\rho})$, the pointwise consistency follows by $\hat{\rho} - \rho = o_P(1)$ and the fact that $y_t = O_P(1)$ for $T_0 + 1 \leq t \leq T$ (due to the stationarity property of u_t).

A.10 Proof of Lemma 7

By assumption, $\max_{K+1 \leq t \leq T} |\hat{P}_t^N - P_t^N| \leq \ell_T \|\hat{\rho} - \rho\|$. Therefore,

$$\frac{1}{T} \sum_{t=K+1}^{T} (\hat{P}_t^N - P_t^N)^2 \leq \ell_T^2 \|\hat{\rho} - \rho\|$$

and

$$\max_{T_0+1 \leq t \leq T} |\hat{P}_t^N - P_t^N| \leq \ell_T \|\hat{\rho} - \rho\|.$$

Since $\ell_T \|\hat{\rho} - \rho\| = o_P(1)$, the desired result follows.

A.11 Proof of Lemma 8

We first derive the following result that is useful in proving Lemma 8.

Lemma 20. *Recall $\varepsilon_t = x_t'\rho + u_t$, where $\rho = (\rho_1, \rho_2, ..., \rho_K)' \in \mathbb{R}^K$ and $x_t = (\varepsilon_{t-1}, \varepsilon_{t-2}, ..., \varepsilon_{t-K})' \in \mathbb{R}^K$. Suppose that the following hold: (1) $\{u_t\}_{t=1}^T$ is an i.i.d sequence with $E(u_1^4)$ uniformly bounded. (2) the roots of $1 - \sum_{j=1}^{K} \rho_j L^j = 0$ are uniformly bounded away from the unit circle.*

Then we have (i) $(T-K)^{-1}\sum_{t=K+1}^{T}u_t^2 = O_P(1)$; (ii) $(T-K)^{-1}\sum_{t=K+1}^{T}x_tu_t = o_P(1)$; (iii) $(T-K)^{-1}\sum_{t=K+1}^{T}\|x_t\|^2 = O_P(1)$. (iv) There exists a constant $\lambda_0 > 0$ such that the smallest eigenvalue of $(T-K)^{-1}\sum_{t=K+1}^{T}x_tx_t'$ is bounded below by λ_0 with probability approaching one.

Proof. **Proof of part (i)**. Part (i) follows by the law of large numbers; see e.g., Theorem 3.1 of White (2014).

Proof of part (ii). Let $\mathcal{F}_t$ be the σ-algebra generated by $\{u_s : s \leq t\}$. First notice that $\{x_tu_t\}_{t=K+1}^{T}$ is a martingale difference sequence with respect to the filtration $\{\mathcal{F}_t\}$. Since ε_t is a stationary process, we have that $E\|x_tu_t\|^2 = \sum_{j=1}^{K}E(\varepsilon_{t-j}^2u_t^2) = \sum_{j=1}^{K}E(\varepsilon_{t-j}^2)E(u_t^2)$ is uniformly bounded bounded. Hence, part (ii) follows by Exercise 3.77 of White (2014).

Proof of part (iii). To see part (iii), notice that $\|x_t\|^2 = x_t'x_t = \sum_{j=1}^{K}\varepsilon_{t-j}^2$. By the analysis on page 215 of Hamilton (1994), for each $1 \leq j \leq K$, $(T-K)^{-1}\sum_{t=K+1}^{T}\varepsilon_{t-j}^2 = E(\varepsilon_{t-j}^2) + o_P(1)$. Thus, part (iii) follows by

$$(T-K)^{-1}\sum_{t=K+1}^{T}\|x_t\|^2 = (T-K)^{-1}\sum_{j=1}^{K}\sum_{t=K+1}^{T}\varepsilon_{t-j}^2 = K\left(E(\varepsilon_t^2) + o_P(1)\right).$$

Proof of part (iv). Similarly, the analysis on page 215 of Hamilton (1994) implies that

$$(T-K)^{-1}\sum_{t=K+1}^{T}x_tx_t' = o_P(1) + Ex_tx_t'.$$

By Proposition 5.1.1 of Brockwell and Davis (2013), $E(x_tx_t')$ has eigenvalues bounded away from zero. Part (iv) follows. □

Now we are ready to prove Lemma 8.

Proof of Lemma 8. Define $\delta_t = \hat{\varepsilon}_t - \varepsilon_t$, $\Delta_t = \hat{x}_t - x_t$, $\tilde{u}_t = u_t + \delta_t - \Delta_t'\rho$ and $a_t = \tilde{u}_t - u_t$. Notice that

$$\hat{\varepsilon}_t = \delta_t + \varepsilon_t = \delta_t + x_t'\rho + u_t = \delta_t + (\hat{x}_t - \Delta_t)'\rho + u_t = \hat{x}_t'\rho + \tilde{u}_t. \tag{55}$$

Therefore,

$$\begin{aligned}\hat{\rho} = \left(\sum_{t=K+1}^{T}\hat{x}_t\hat{x}_t'\right)^{-1}\left(\sum_{t=K+1}^{T}\hat{x}_t\hat{\varepsilon}_t\right) &= \left(\sum_{t=K+1}^{T}\hat{x}_t\hat{x}_t'\right)^{-1}\left(\sum_{t=K+1}^{T}\hat{x}_t(\hat{x}_t'\rho + \tilde{u}_t)\right)\\ &= \rho + \left(\sum_{t=K+1}^{T}\hat{x}_t\hat{x}_t'\right)^{-1}\left(\sum_{t=K+1}^{T}\hat{x}_t\tilde{u}_t\right).\end{aligned} \tag{56}$$

The rest of the proof proceeds in three steps. First two steps show that $(T-K)^{-1}\sum_{t=K+1}^{T}\hat{x}_t\hat{x}_t'$ is well-behaved and $(T-K)^{-1}\sum_{t=K+1}^{T}\hat{x}_t\tilde{u}_t = o_P(1)$. This would imply $\hat{\rho} = \rho + o_P(1)$. In the third step, we derive the final result.

Step 1: show that $\left[(T-K)^{-1}\sum_{t=K+1}^{T}\hat{x}_t\hat{x}_t'\right]^{-1} = O_P(1)$.

It is not hard to see that $\|\Delta_t\|^2 = \sum_{s=t-1}^{t-K} \delta_s^2$. Therefore,

$$\sum_{t=K+1}^{T} \|\Delta_t\|^2 = \sum_{t=K+1}^{T} \sum_{s=t-1}^{t-K} \delta_s^2 \leq K \sum_{t=1}^{T} \delta_t^2 \overset{\text{(i)}}{=} o_P(T), \tag{57}$$

where (i) follows by the assumption of $T^{-1}\sum_{t=1}^{T} \delta_t^2 = o_P(1)$. Notice that

$$\begin{aligned}\left\| \sum_{t=K+1}^{T} (\hat{x}_t \hat{x}_t' - x_t x_t') \right\| &= \left\| \sum_{t=K+1}^{T} (x_t \Delta_t' + \Delta_t x_t' + \Delta_t \Delta_t') \right\| \\ &\leq 2 \sum_{t=K+1}^{T} \|x_t\| \cdot \|\Delta_t\| + \sum_{t=K+1}^{T} \|\Delta_t\|^2 \\ &\leq 2\sqrt{\left(\sum_{t=K+1}^{T} \|x_t\|^2 \right) \left(\sum_{t=K+1}^{T} \|\Delta_t\|^2 \right)} + \sum_{t=K+1}^{T} \|\Delta_t\|^2 \overset{\text{(i)}}{=} o_P(T), \end{aligned} \tag{58}$$

where (i) follows by (57) and Lemma 20. Thus,

$$\left\| \frac{1}{T-K} \sum_{t=K+1}^{T} (\hat{x}_t \hat{x}_t' - x_t x_t') \right\| = o_P(1).$$

By Lemma 20, the smallest eigenvalue of $(T-K)^{-1}\sum_{t=K+1}^{T} x_t x_t'$ is bounded below by a positive constant with probability approaching one. It follows that

$$\left[(T-K)^{-1} \sum_{t=K+1}^{T} \hat{x}_t \hat{x}_t' \right]^{-1} = O_P(1). \tag{59}$$

Step 2: show that $(T-K)^{-1}\sum_{t=K+1}^{T} \hat{x}_t \tilde{u}_t = o_P(1)$.
By Lemma 20, we have

$$(T-K)^{-1} \sum_{t=K+1}^{T} x_t u_t = o_P(1). \tag{60}$$

Notice that

$$\begin{aligned}\left\| \frac{1}{T-K} \sum_{t=K+1}^{T} (\hat{x}_t \tilde{u}_t - x_t u_t) \right\| &= \left\| \frac{1}{T-K} \sum_{t=K+1}^{T} (\Delta_t u_t + x_t a_t + \Delta_t a_t) \right\| \\ &\leq \frac{1}{T-K} \sum_{t=K+1}^{T} (\|\Delta_t u_t\| + \|x_t a_t\| + \|\Delta_t a_t\|) \end{aligned}$$

$$
\begin{aligned}
&\leq \sqrt{\left(\frac{1}{T-K}\sum_{t=K+1}^{T}\|\Delta_t\|^2\right)\left(\frac{1}{T-K}\sum_{t=K+1}^{T}u_t^2\right)} \\
&\quad + \sqrt{\left(\frac{1}{T-K}\sum_{t=K+1}^{T}\|x_t\|^2\right)\left(\frac{1}{T-K}\sum_{t=K+1}^{T}a_t^2\right)} \\
&\quad + \sqrt{\left(\frac{1}{T-K}\sum_{t=K+1}^{T}\|\Delta_t\|^2\right)\left(\frac{1}{T-K}\sum_{t=K+1}^{T}a_t^2\right)}. \qquad (61)
\end{aligned}
$$

We observe that

$$
\begin{aligned}
\sum_{t=K+1}^{T} a_t^2 = \sum_{t=K+1}^{T} (\delta_t - \Delta_t'\rho)^2 &\leq 2\sum_{t=K+1}^{T}\delta_t^2 + 2\sum_{t=K+1}^{T}(\Delta_t'\rho)^2 \\
&\leq 2\sum_{t=1}^{T}\delta_t^2 + 2\|\rho\|^2\sum_{t=K+1}^{T}\|\Delta_t\|^2 \overset{\text{(i)}}{=} O_P(T), \qquad (62)
\end{aligned}
$$

where (i) follows by (57) and the assumption of $T^{-1}\sum_{t=1}^{T}\delta_t^2 = o_P(1)$. Combining (61) with (57) and (62), we obtain

$$
\begin{aligned}
&\left\|\frac{1}{T-K}\sum_{t=K+1}^{T}(\hat{x}_t\tilde{u}_t - x_t u_t)\right\| \\
&\leq \sqrt{o_P(1)\left(\frac{1}{T-K}\sum_{t=K+1}^{T}u_t^2\right)} + \sqrt{\left(\frac{1}{T-K}\sum_{t=K+1}^{T}\|x_t\|^2\right)o_P(1)} + \sqrt{o_P(1)\times o_P(1)} \overset{\text{(i)}}{=} o_P(1), \qquad (63)
\end{aligned}
$$

where (i) follows by Lemma 20. Now we combine (60) and (63), obtaining

$$
(T-K)^{-1}\sum_{t=K+1}^{T}\hat{x}_t\tilde{u}_t = o_P(1). \qquad (64)
$$

By (56) together with (59) and (64), it follows that

$$
\hat{\rho} - \rho = o_P(1). \qquad (65)
$$

Step 3: show the desired result.
Recall that $\hat{u}_t = \hat{\varepsilon}_t - \hat{x}_t'\hat{\rho}$. Hence,

$$
\hat{u}_t - u_t = (\hat{\varepsilon}_t - \hat{x}_t'\hat{\rho}) - u_t \overset{\text{(i)}}{=} (\hat{x}_t'(\rho - \hat{\rho}) + \tilde{u}_t) - u_t = \hat{x}_t'(\rho - \hat{\rho}) + a_t, \qquad (66)
$$

where (i) follows by (55). Therefore, we have

$$
\begin{aligned}
\sum_{t=K+1}^{T} (\hat{u}_t - u_t)^2 &= \sum_{t=K+1}^{T} (\hat{x}_t'(\rho - \hat{\rho}) + a_t)^2 \\
&\leq 2 \sum_{t=K+1}^{T} (\hat{x}_t'(\hat{\rho} - \rho))^2 + 2 \sum_{t=K+1}^{T} a_t^2 \\
&\leq 2\|\hat{\rho} - \rho\|^2 \sum_{t=K+1}^{T} \|\hat{x}_t\|^2 + 2 \sum_{t=K+1}^{T} a_t^2 \\
&= 2\|\hat{\rho} - \rho\|^2 \left(\sum_{t=K+1}^{T} \operatorname{trace}(x_t x_t') + \sum_{t=K+1}^{T} \operatorname{trace}(\hat{x}_t \hat{x}_t' - x_t x_t') \right) + 2 \sum_{t=K+1}^{T} a_t^2 \\
&\overset{\text{(i)}}{\leq} o_P(1) \times (O_P(T) + o_P(T)) + o_P(T) = o_P(T),
\end{aligned}
$$

where (i) follows by (58), (65), (62) and Lemma 20.

To see the pointwise result, we notice that by (66) and (65), it suffices to verify that $a_t = o_P(1)$ and $\hat{x}_t = O_P(1)$ for $T_0 + 1 \leq t \leq T$.

Since $\hat{x}_t - x_t = (\delta_{t-1}, \delta_{t-2}, ..., \delta_{t-K})'$, the assumption of pointwise convergence of $\hat{\varepsilon}_t$ (i.e., $\delta_t = o_P(1)$ for $T_0 + 1 - K \leq t \leq T$) implies that $\hat{x}_t - x_t = o_P(1)$ for $T_0 + 1 \leq t \leq T$. Since $x_t = O_P(1)$ due to the stationarity condition, we have $\hat{x}_t = O_P(1)$ for $T_0 + 1 \leq t \leq T$.

Since both δ_t and Δ_t are both $o_P(1)$ for $T_0 + 1 \leq t \leq T$, so is $a_t = \delta_t - \Delta_t'\rho$. Hence, we have proved the pointwise result. The proof is complete. □

A.12 Proof of Lemma 9

Fix an arbitrary $\eta > 0$. Define $a_\eta = \inf_{\|\beta - \beta_*\|_2 \geq \eta}(L(\beta) - L(\beta_*))/3$. By the compactness of $\mathcal{B}$ and the uniqueness of the minimum of $L(\cdot)$, we have $a_\eta > 0$.

(Otherwise, one can find a sequence $\{\beta_k\}_{k=1}^{\infty}$ with $\|\beta_k - \beta_*\|_2 \geq \eta$ for all $k \geq 1$ with $L(\beta_k) \to L(\beta_*)$. By compactness of $\mathcal{B}$ implies that some subsequence of β_k converges to a point $\beta_{**} \in \mathcal{B}$. Clearly, $\|\beta_{**} - \beta_*\|_2 \geq \eta$. The continuity of $L(\cdot)$ implies $L(\beta_{**}) = L(\beta_*)$. This contradicts the uniqueness of the minimum of $L(\cdot)$.)

Define the event

$$
\mathcal{M} = \left\{ \sup_{\beta} |\hat{L}(\mathbf{Z}; \beta) - L(\beta)| \leq a_\eta \right\} \bigcap \left\{ \max_{H \in \mathcal{H}} \sup_{\beta} |\hat{L}(\mathbf{Z}_H; \beta) - L(\beta)| \leq a_\eta \right\}.
$$

By the assumption, we know $P(\mathcal{M}) = 1 - o(1)$.

Notice that on the event $\mathcal{M}$,

$$
L(\hat{\beta}(\mathbf{Z})) - L(\beta_*) \leq a_\eta + \hat{L}(\mathbf{Z}; \hat{\beta}(\mathbf{Z})) - L(\beta_*) \leq 2a_\eta + \hat{L}(\mathbf{Z}; \hat{\beta}(\mathbf{Z})) - \hat{L}(\mathbf{Z}; \beta_*) \leq 2a_\eta.
$$

It follows by the definition of a_η that $\|\hat{\beta}(\mathbf{Z}) - \beta_*\|_2 \leq \eta$ on the event $\mathcal{M}$. Thus, $P(\|\hat{\beta} - \beta_*\|_2 \leq \eta) \geq P(\mathcal{M}) = 1 - o(1)$. Since $\eta > 0$ is arbitrary, we have $\|\hat{\beta}(\mathbf{Z}) - \beta_*\|_2 = o_P(1)$.

By the same analysis, we have that on the event $\mathcal{M}$, $\|\hat{\beta}(\mathbf{Z}_H) - \beta_*\|_2 \leq \eta$ for all $H \in \mathcal{H}$. Thus, on the event $\mathcal{M}$, $\max_{H\in\mathcal{H}} \|\hat{\beta}(\mathbf{Z}_H) - \beta_*\|_2 \leq \eta$. We have that $\max_{H\in\mathcal{H}} \|\hat{\beta}(\mathbf{Z}_H) - \beta_*\|_2 = o_P(1)$. The desired result follows.

A.13 Proof of Lemma 10

For notational simplicity, we write $\hat{\beta} = \hat{\beta}(\mathbf{Z})$ and $\hat{\beta}_H = \hat{\beta}(\mathbf{Z}_H)$. Define the event $\mathcal{M} = \mathcal{M}_1 \bigcap \mathcal{M}_2$, where $\mathcal{M}_1 = \{\min_{\|v\|_0\leq m} v'\hat{\Sigma}v/\|v\|_2^2 \geq \kappa_0\} \bigcap \{\|\hat{\beta}\|_0 \leq s/2\}$ and

$$\mathcal{M}_2 = \bigcap_{H\in\mathcal{H}} \left(\left\{ \|\hat{\Sigma}_H - \hat{\Sigma}\|_\infty \leq c_T,\ \|\hat{\mu}_H - \hat{\mu}\|_\infty \leq c_T \right\} \bigcap \left\{ \max_{H\in\mathcal{H}} \|\hat{\beta}_H\|_0 \leq s/2 \right\} \right).$$

By assumption, $P(\mathcal{M}) \geq 1 - \gamma_{1,T} - \gamma_{2,T} - \gamma_{3,T}$. The rest of the argument are statements on the event $\mathcal{M}$.

Fix $H \in \mathcal{H}$, let $\Delta = \hat{\beta}_H - \hat{\beta}$. Define $\xi = \hat{\mu} - \hat{\Sigma}\hat{\beta}$ and $\xi_H = \hat{\mu}_H - \hat{\Sigma}_H\hat{\beta}$. Since $\|\hat{\beta}\|_1 \leq K$, we have

$$\|\xi_H - \xi\|_\infty \leq \|\hat{\mu}_H - \hat{\mu}\|_\infty + \|\hat{\Sigma}_H - \hat{\Sigma}\|_\infty \|\hat{\beta}\|_1 \leq c_T(1 + K). \tag{67}$$

When $\Delta = 0$, the result clearly holds. Now we consider the case with $\Delta \neq 0$.

Step 1: show that on the event $\mathcal{M}$, $0 \leq \lambda^2\Delta'\hat{\Sigma}\Delta - \lambda\Delta'\hat{\mu} \leq c_T K^2 + 2c_T K$ for any $\lambda \in [0, 1]$.

Recall that $\hat{Q}(\beta) = \beta'\hat{\Sigma}\beta - 2\hat{\mu}'\beta + T^{-1}\sum_{t=1}^T Y_t^2$. Since the term $T^{-1}\sum_{t=1}^T Y_t^2$ does not affect the minimizer, we modify $\hat{Q}$ by dropping this term. With a slight abuse of notation, we still use the symbol $\hat{Q}(\beta) = \beta'\hat{\Sigma}\beta - 2\hat{\mu}'\beta$ and $\hat{Q}_H(\beta) = \beta'\hat{\Sigma}_H\beta - 2\hat{\mu}'_H\beta$. Therefore,

$$\hat{Q}(\beta) - \hat{Q}_H(\beta) = \beta'(\hat{\Sigma} - \hat{\Sigma}_H)\beta - 2(\hat{\mu} - \hat{\mu}_H)'\beta.$$

Since $\sup_{\beta\in\mathcal{W}} \|\beta\|_1 \leq K$, we have that on the event $\mathcal{M}$,

$$\sup_{\beta\in\mathcal{W}} \left| \hat{Q}(\beta) - \hat{Q}_H(\beta) \right| \leq c_T K^2 + 2c_T K.$$

Let $\bar{\beta} = \hat{\beta} + \lambda\Delta$, where $\lambda \in [0, 1]$. Then clearly, $\bar{\beta} = \lambda\hat{\beta}_H + (1 - \lambda)\hat{\beta}$. By definition of $\hat{\beta}$, we have that $\hat{Q}(\hat{\beta}) \leq \hat{Q}(\bar{\beta})$, which means that

$$\lambda^2\Delta'\hat{\Sigma}\Delta - 2\lambda\xi'\Delta \geq 0. \tag{68}$$

Clearly, $\hat{Q}_H(\hat{\beta}_H) \leq \hat{Q}_H(\bar{\beta})$ for any $\lambda \in [0, 1]$. Hence, $\hat{Q}_H(\hat{\beta} + \Delta) \leq \hat{Q}_H(\hat{\beta} + \lambda\Delta)$ for any $\lambda \in (0, 1)$. By $\hat{Q}_H(\beta) = \beta'\hat{\Sigma}_H\beta - 2\hat{\mu}'_H\beta$, this simplifies to $(1 + \lambda)\Delta'\hat{\Sigma}_H\Delta \leq 2\xi'_H\Delta$. Since λ can be arbitrarily close to one, this means $\Delta'\hat{\Sigma}_H\Delta \leq \xi'_H\Delta$. It follows that for any $\lambda \in [0, 1]$, we have

$$\lambda^2\Delta'\hat{\Sigma}_H\Delta \leq 2\lambda\xi'_H\Delta.$$

Notice that

$$0 \leq 2\lambda\xi'_H\Delta - \lambda^2\Delta'\hat{\Sigma}_H\Delta = 2\lambda\xi'\Delta - \lambda^2\Delta'\hat{\Sigma}\Delta + 2\lambda(\xi_H - \xi)'\Delta + \lambda^2\Delta'(\hat{\Sigma} - \hat{\Sigma}_H)\Delta$$

$$\leq 2\lambda\xi'\Delta - \lambda^2\Delta'\hat{\Sigma}\Delta + 2\lambda\|\xi_H - \xi\|_\infty\|\Delta\|_1 + \lambda^2\|\hat{\Sigma} - \hat{\Sigma}_H\|_\infty\|\Delta\|_1^2.$$

It follows, by (67) and $\|\Delta\|_1 \leq \|\hat{\beta}_H\|_1 + \|\hat{\beta}\|_1 \leq 2K$, that

$$\lambda^2\Delta'\Sigma\Delta - 2\lambda\xi'\Delta \leq 2\lambda\|\xi_H - \xi\|_\infty\|\Delta\|_1 + \lambda^2\|\hat{\Sigma} - \hat{\Sigma}_H\|_\infty\|\Delta\|_1^2 \leq 4c_T(1+K)K + 4c_TK^2. \quad (69)$$

Since (68) and (69) hold for any $\lambda \in [0,1]$, we have that

$$0 \leq \lambda^2\Delta'\hat{\Sigma}\Delta - 2\lambda\xi'\Delta \leq 4c_TK(2K+1) \qquad \forall\lambda \in [0,1]. \quad (70)$$

Step 2: show the desired result.

Since $0 \leq \lambda^2\Delta'\hat{\Sigma}\Delta - 2\lambda\xi'\Delta$ for any $\lambda \in (0,1)$, we have that $\xi'\Delta \leq \lambda\Delta'\hat{\Sigma}\Delta/2$ for any $\lambda \in (0,1)$. Thus,

$$\xi'\Delta \leq 0.$$

Hence, by the second inequality in (70), for any $\lambda \in [0,1]$,

$$\lambda^2\Delta'\hat{\Sigma}\Delta \leq \lambda^2\Delta'\hat{\Sigma}\Delta - 2\lambda\xi'\Delta \leq 4c_TK(2K+1).$$

Now we take $\lambda = 1$, which implies

$$\Delta'\hat{\Sigma}\Delta \leq 4c_TK(2K+1).$$

Since $\|\Delta\|_0 \leq \|\hat{\beta}_H\|_0 + \|\hat{\beta}\|_0 \leq s$ and $\|\Delta\|_1 \leq \sqrt{\|\Delta\|_0}\|\Delta\|_2$, it follows that

$$4c_TK(2K+1) \geq \Delta'\hat{\Sigma}\Delta \geq \kappa_1\|\Delta\|_2^2 \geq \kappa_1 s^{-1}\|\Delta\|_1^2.$$

Hence, $\|\Delta\|_1 \leq 2\sqrt{\kappa_1 s c_T K(2K+1)}$ and

$$\left|(Y_t - X_t'\hat{\beta}) - (Y_t - X_t'\hat{\beta}_H)\right| = |X_t'\Delta| \leq \|X_t\|_\infty\|\Delta\|_1 \leq 2\kappa_2\sqrt{\kappa_1 s c_T K(2K+1)}.$$

On the event $\mathcal{M}$, the above bound holds for all $H \in \mathcal{H}$. The desired result follows by $P(\mathcal{M}) \geq 1 - \gamma_{1,T} - \gamma_{2,T} - \gamma_{3,T}$.

B Tables and Figures

Table 1: Size i.i.d. Data

	DGP1								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
20	0.09	0.10	0.09	0.10	0.10	0.09	0.09	0.10	0.09
50	0.10	0.10	0.10	0.09	0.10	0.09	0.09	0.10	0.10
100	0.10	0.11	0.10	0.10	0.10	0.10	0.10	0.11	0.10

	DGP2								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
20	0.09	0.09	0.10	0.09	0.10	0.10	0.09	0.09	0.09
50	0.10	0.10	0.10	0.10	0.10	0.11	0.10	0.10	0.11
100	0.09	0.10	0.11	0.09	0.10	0.10	0.10	0.10	0.10

	DGP3								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
20	0.10	0.10	0.10	0.10	0.10	0.09	0.09	0.10	0.09
50	0.10	0.10	0.10	0.10	0.09	0.09	0.10	0.09	0.10
100	0.10	0.10	0.10	0.09	0.09	0.10	0.11	0.10	0.10

	DGP4								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
20	0.09	0.09	0.10	0.10	0.09	0.10	0.10	0.09	0.10
50	0.10	0.10	0.11	0.10	0.10	0.10	0.10	0.10	0.10
100	0.10	0.10	0.09	0.10	0.10	0.09	0.10	0.10	0.09

Notes: Simulation design as described in the main text with $\rho_\epsilon = \rho_u = 0$. Nominal level $= 0.1$. Based on simulations with 5000 repetitions.

Table 2: Size Dependent Data

	DGP1								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
20	0.13	0.13	0.13	0.12	0.11	0.11	0.11	0.11	0.11
50	0.12	0.11	0.11	0.12	0.12	0.12	0.12	0.12	0.12
100	0.10	0.11	0.11	0.11	0.12	0.12	0.11	0.12	0.11

	DGP2								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
20	0.12	0.12	0.12	0.11	0.12	0.12	0.12	0.12	0.11
50	0.11	0.11	0.11	0.11	0.11	0.12	0.11	0.12	0.13
100	0.10	0.10	0.10	0.11	0.10	0.11	0.11	0.11	0.11

	DGP3								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
20	0.10	0.09	0.11	0.10	0.09	0.09	0.13	0.11	0.11
50	0.10	0.10	0.10	0.10	0.10	0.10	0.12	0.12	0.12
100	0.10	0.10	0.10	0.10	0.10	0.10	0.11	0.12	0.12

	DGP4								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
20	0.11	0.11	0.11	0.11	0.10	0.09	0.11	0.10	0.10
50	0.10	0.10	0.11	0.11	0.11	0.11	0.11	0.11	0.11
100	0.10	0.10	0.11	0.10	0.11	0.11	0.10	0.11	0.11

Notes: Simulation design as described in the main text with $\rho_\epsilon = \rho_u = 0.6$. Nominal level $= 0.1$. Based on simulations with 5000 repetitions.

Table 3: Power i.i.d. Data

	DGP1								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
20	0.57	0.58	0.58	0.53	0.52	0.53	0.49	0.50	0.49
50	0.61	0.61	0.60	0.57	0.57	0.57	0.56	0.56	0.55
100	0.61	0.63	0.62	0.59	0.60	0.59	0.58	0.59	0.58

	DGP2								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
20	0.46	0.44	0.43	0.47	0.46	0.42	0.48	0.47	0.45
50	0.49	0.48	0.45	0.56	0.54	0.51	0.57	0.55	0.53
100	0.53	0.49	0.47	0.60	0.59	0.57	0.60	0.61	0.59

	DGP3								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
20	0.18	0.18	0.17	0.10	0.11	0.11	0.49	0.49	0.48
50	0.20	0.20	0.19	0.13	0.12	0.13	0.57	0.56	0.56
100	0.20	0.20	0.21	0.12	0.13	0.13	0.60	0.60	0.59

	DGP4								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
20	0.32	0.34	0.33	0.39	0.40	0.39	0.30	0.32	0.31
50	0.34	0.36	0.35	0.40	0.41	0.42	0.33	0.35	0.35
100	0.37	0.37	0.37	0.44	0.43	0.44	0.37	0.37	0.38

Notes: Simulation design as described in the main text with $\rho_\epsilon = \rho_u = 0$. Nominal level $= 0.1$. Based on simulations with 5000 repetitions.

Table 4: Power Dependent Data

	DGP1								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
20	0.63	0.65	0.64	0.61	0.63	0.65	0.60	0.62	0.62
50	0.65	0.64	0.64	0.65	0.65	0.67	0.65	0.64	0.67
100	0.65	0.62	0.64	0.64	0.64	0.66	0.64	0.64	0.66
	DGP2								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
20	0.53	0.50	0.50	0.55	0.53	0.53	0.58	0.56	0.57
50	0.54	0.51	0.48	0.60	0.59	0.56	0.64	0.62	0.60
100	0.53	0.50	0.48	0.61	0.59	0.59	0.64	0.62	0.62
	DGP3								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
20	0.18	0.19	0.20	0.12	0.12	0.14	0.61	0.61	0.62
50	0.19	0.20	0.20	0.12	0.13	0.14	0.64	0.65	0.66
100	0.19	0.19	0.20	0.12	0.14	0.15	0.64	0.64	0.66
	DGP4								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
20	0.34	0.34	0.36	0.40	0.41	0.43	0.34	0.33	0.35
50	0.36	0.36	0.38	0.43	0.43	0.45	0.36	0.38	0.40
100	0.36	0.37	0.36	0.43	0.44	0.43	0.37	0.38	0.38

Notes: Simulation design as described in the main text with $\rho_\epsilon = \rho_u = 0.6$. Nominal level $= 0.1$. Based on simulations with 5000 repetitions.

Table 5: Placebo Specification Tests

	Rape rate					
	Moving Block Permutations			i.i.d. Permutations		
Periods	Diff-in-Diffs	Synth. Control	Constr. Lasso	Diff-in-Diffs	Synth. Control	Constr. Lasso
2003	0.46	0.74	0.59	0.47	0.75	0.60
2002 – 2003	0.74	0.51	0.21	0.81	0.56	0.18
2001 – 2003	0.44	0.36	0.44	0.58	0.31	0.38
	Log female gonorrhea					
	Moving Block Permutations			i.i.d. Permutations		
Periods	Diff-in-Diffs	Synth. Control	Constr. Lasso	Diff-in-Diffs	Synth. Control	Constr. Lasso
2003	0.37	0.42	0.84	0.37	0.42	0.83
2002 – 2003	0.53	0.63	1.00	0.55	0.61	0.96
2001 – 2003	0.58	0.74	0.95	0.65	0.81	0.98

Table 6: Zero effect null hypothesis

Rape rate					
Moving Block Permutations			i.i.d. Permutations		
Diff-in-Diffs	Synth. Control	Constr. Lasso	Diff-in-Diffs	Synth. Control	Constr. Lasso
0.04	0.02	0.02	0.00	0.00	0.00
Log female gonorrhea					
Moving Block Permutations			i.i.d. Permutations		
Diff-in-Diffs	Synth. Control	Constr. Lasso	Diff-in-Diffs	Synth. Control	Constr. Lasso
0.04	0.04	0.04	0.00	0.00	0.00

Figure 2: Power Curves DGP1

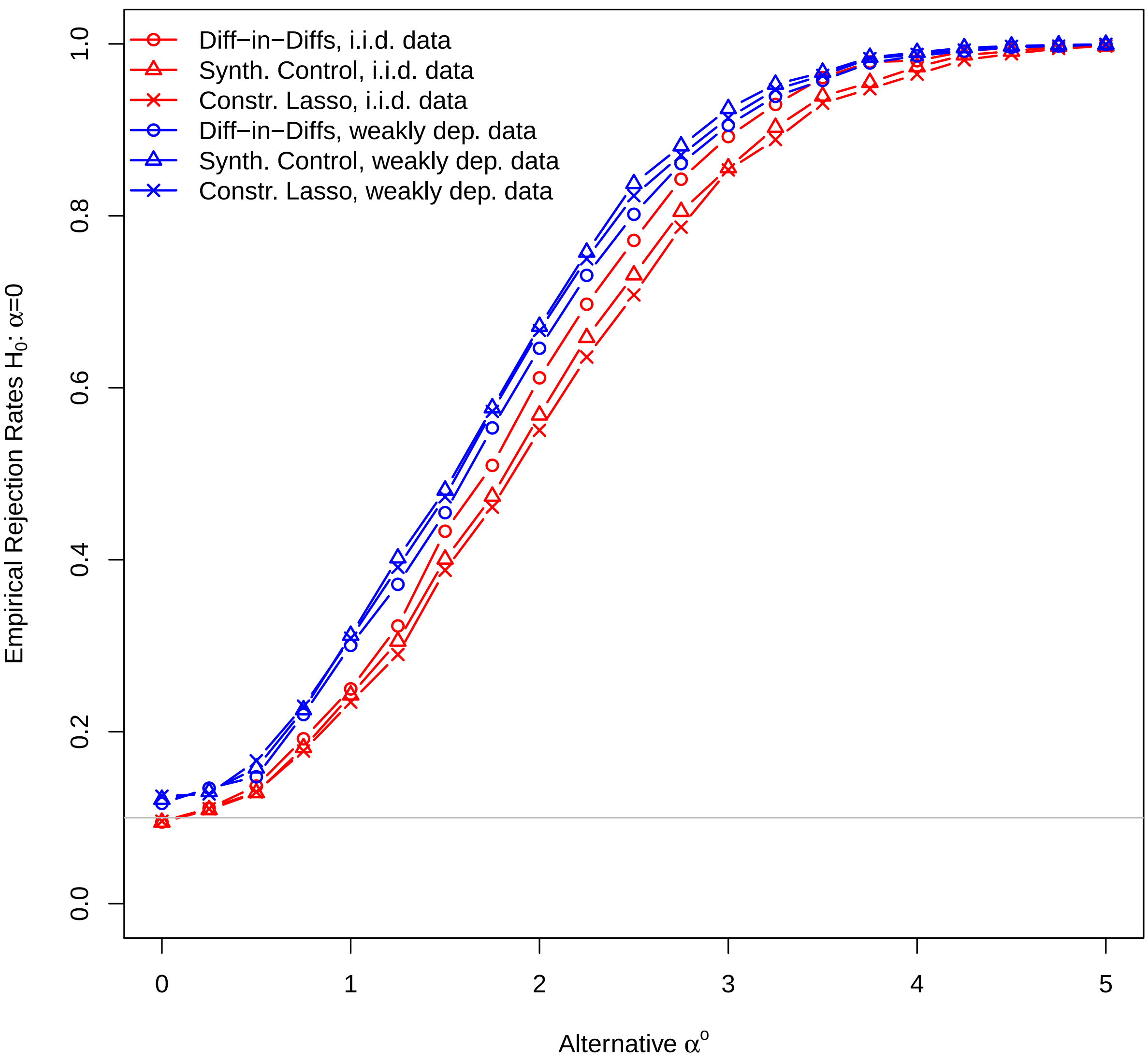


Figure 3: Power Curves DGP2

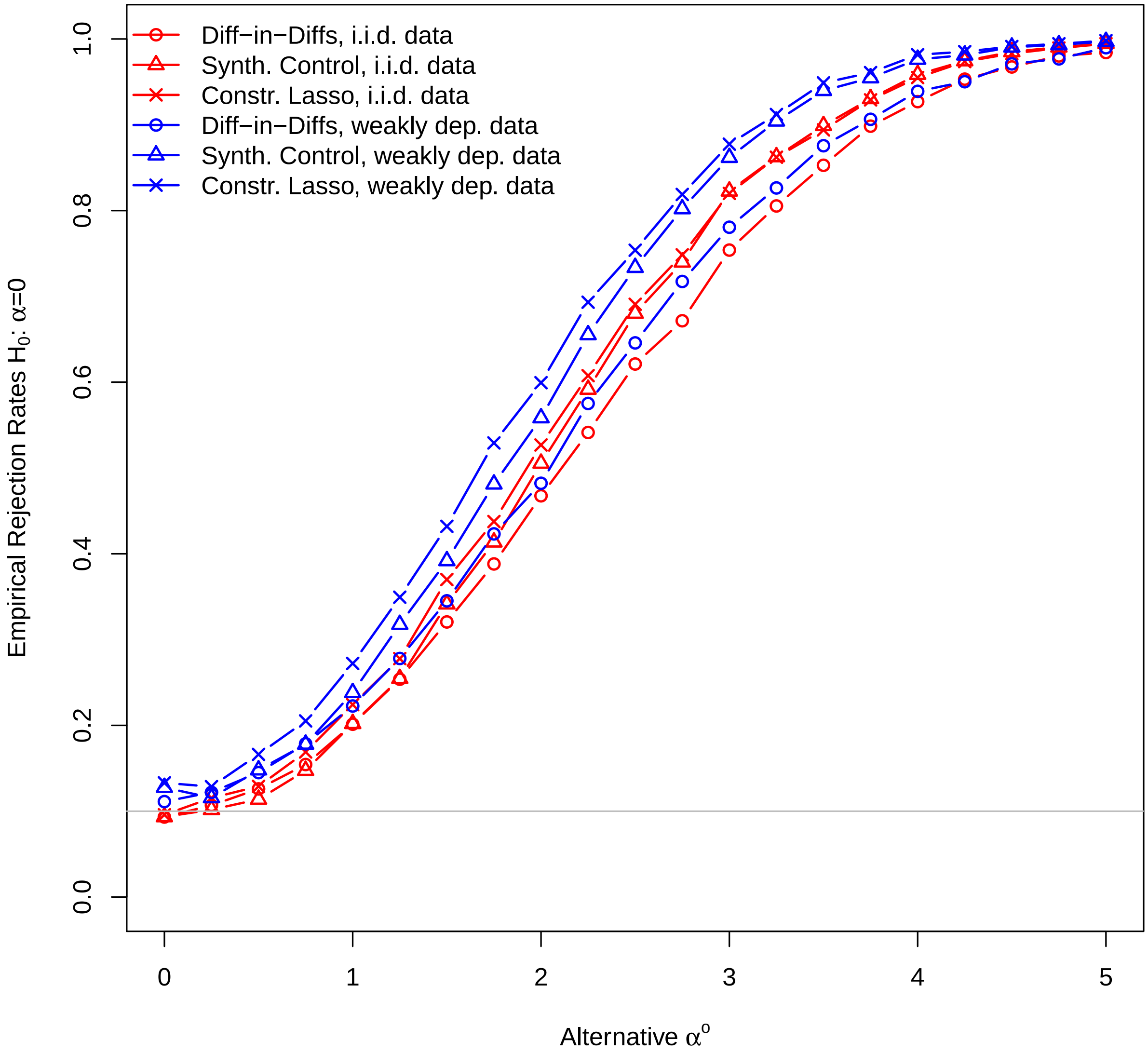


Figure 4: Power Curves DGP3

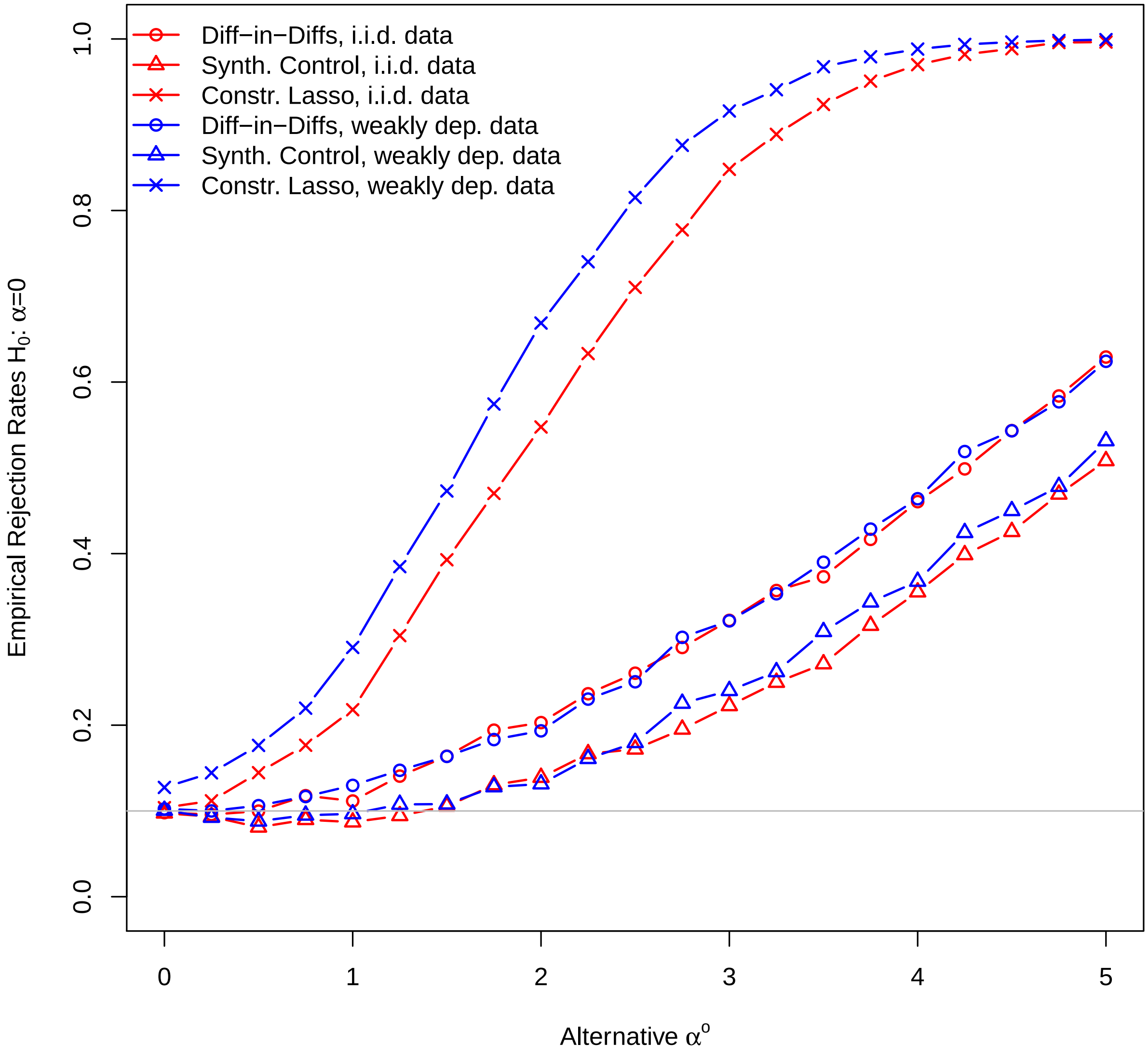


Figure 5: Power Curves DGP4

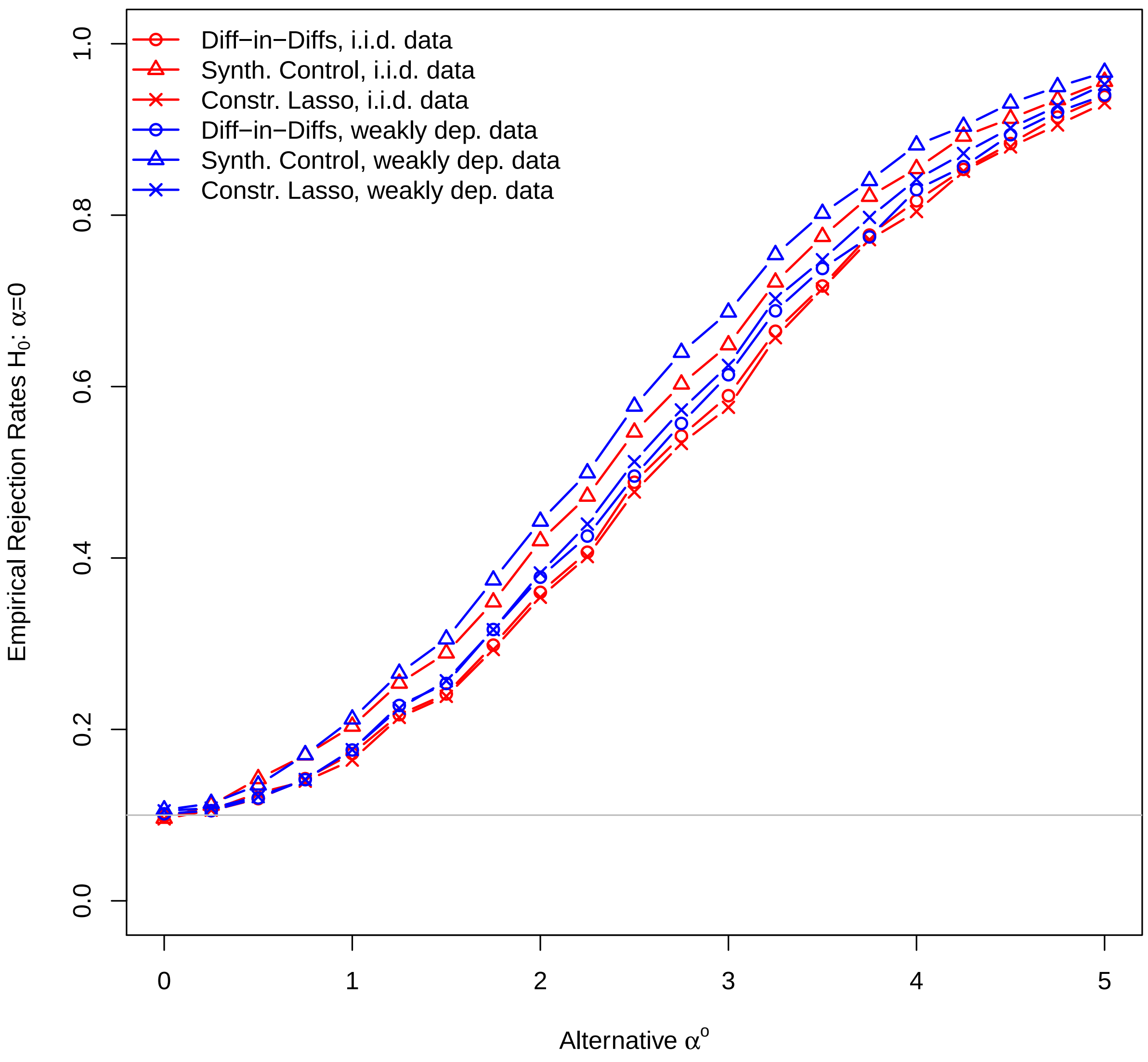


Figure 6: Raw Data

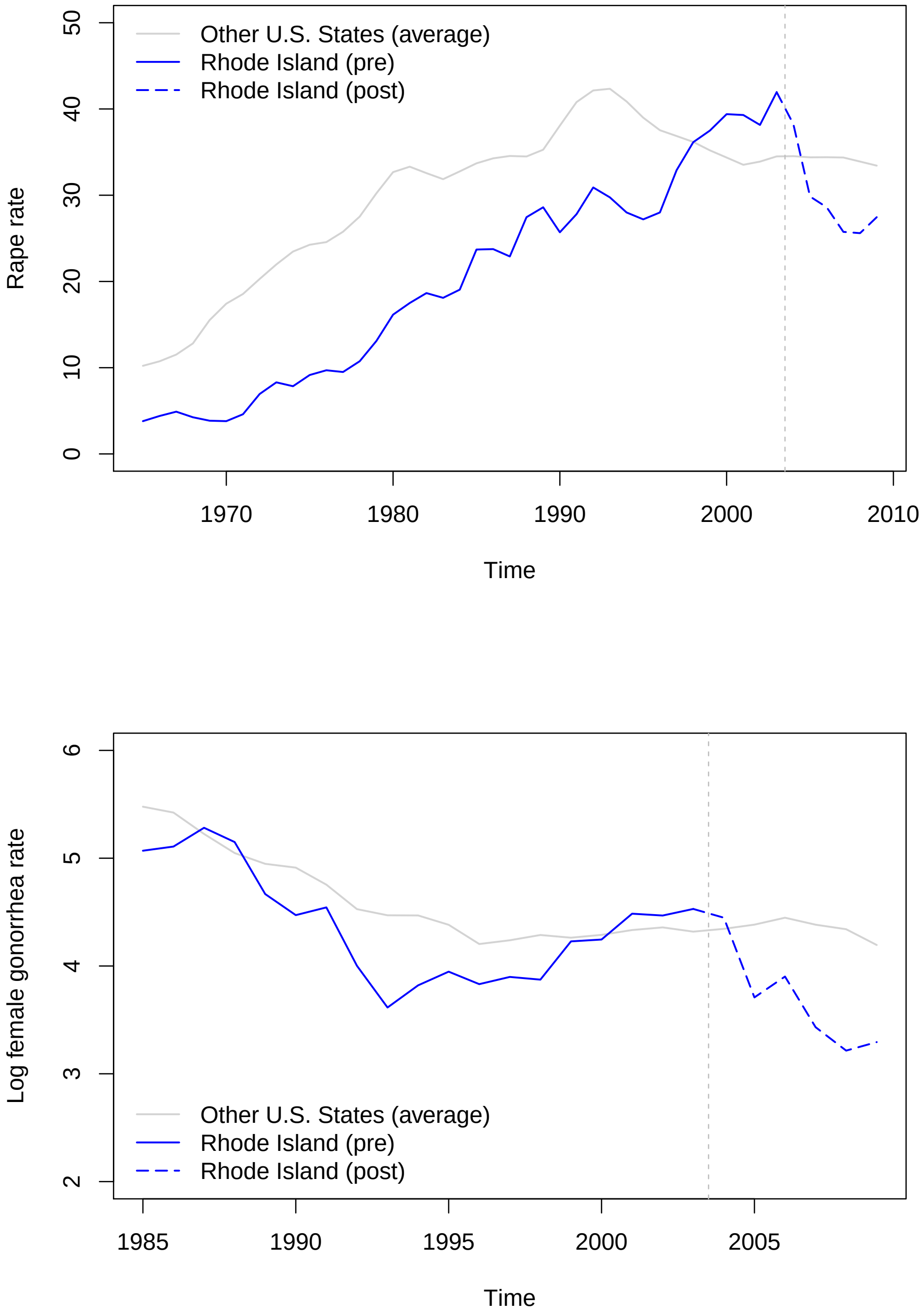


Figure 7: Histograms Placebo Tests Rape Rate

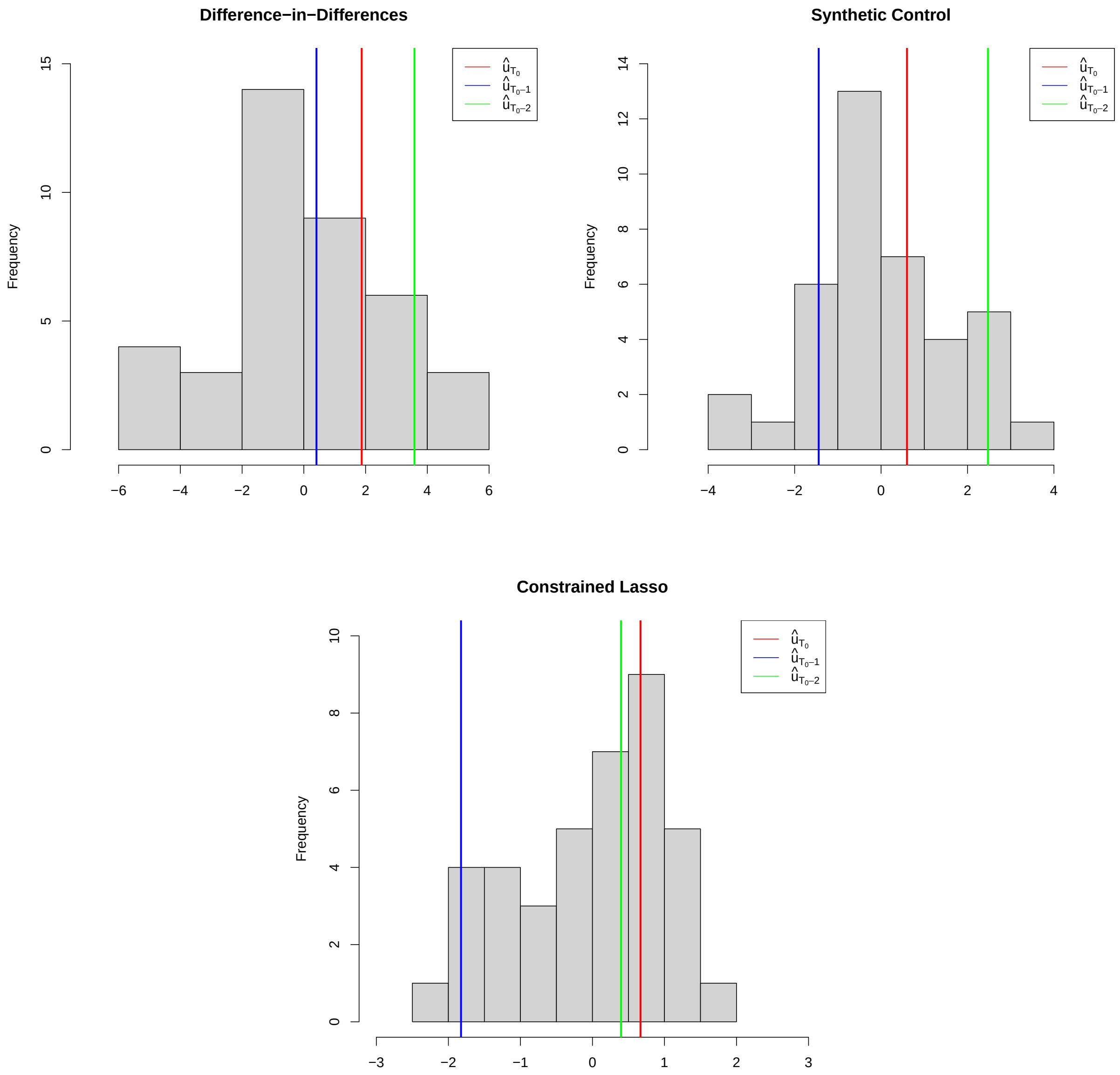


Figure 8: Histograms Placebo Tests Gonorrhea

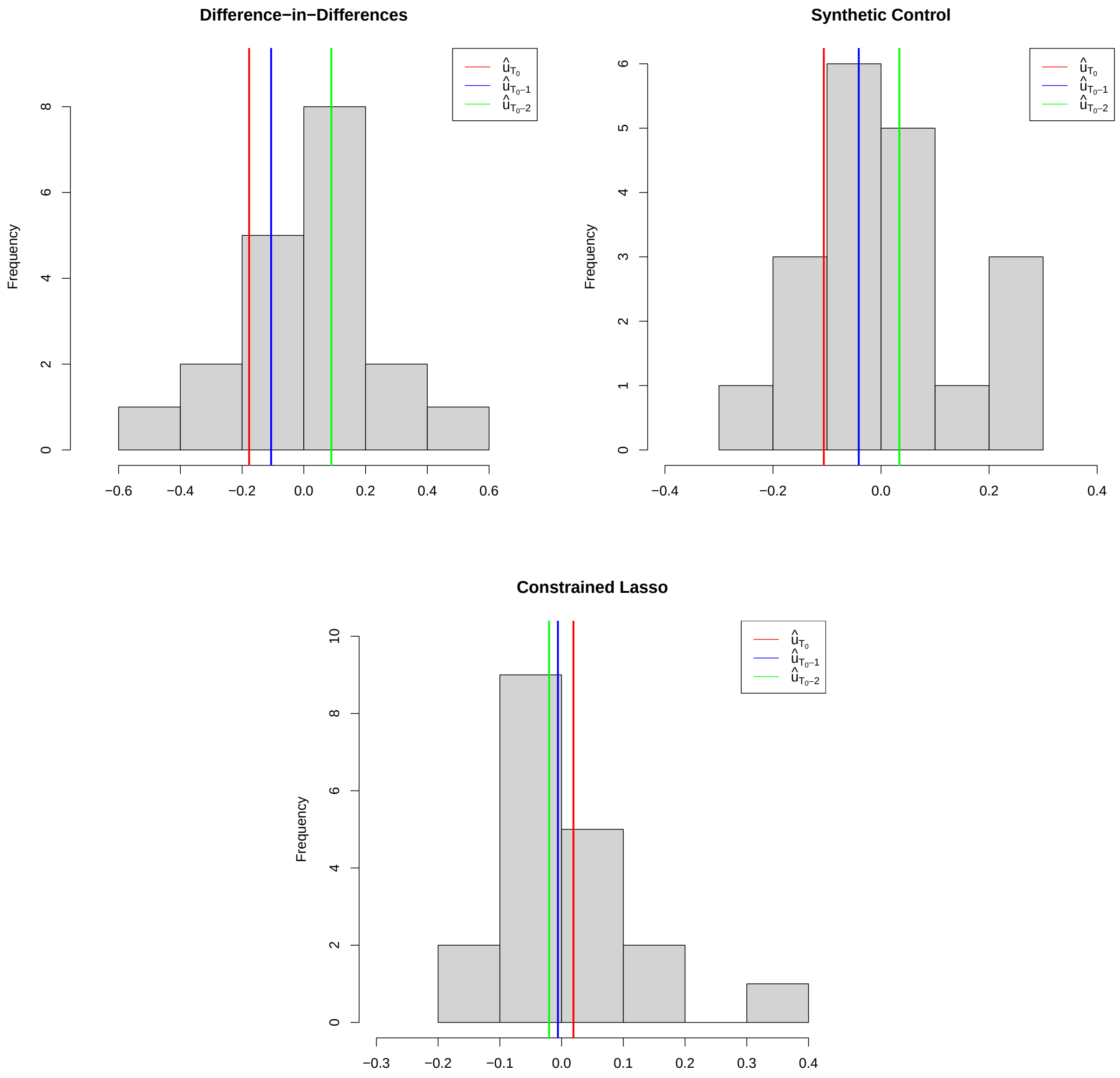
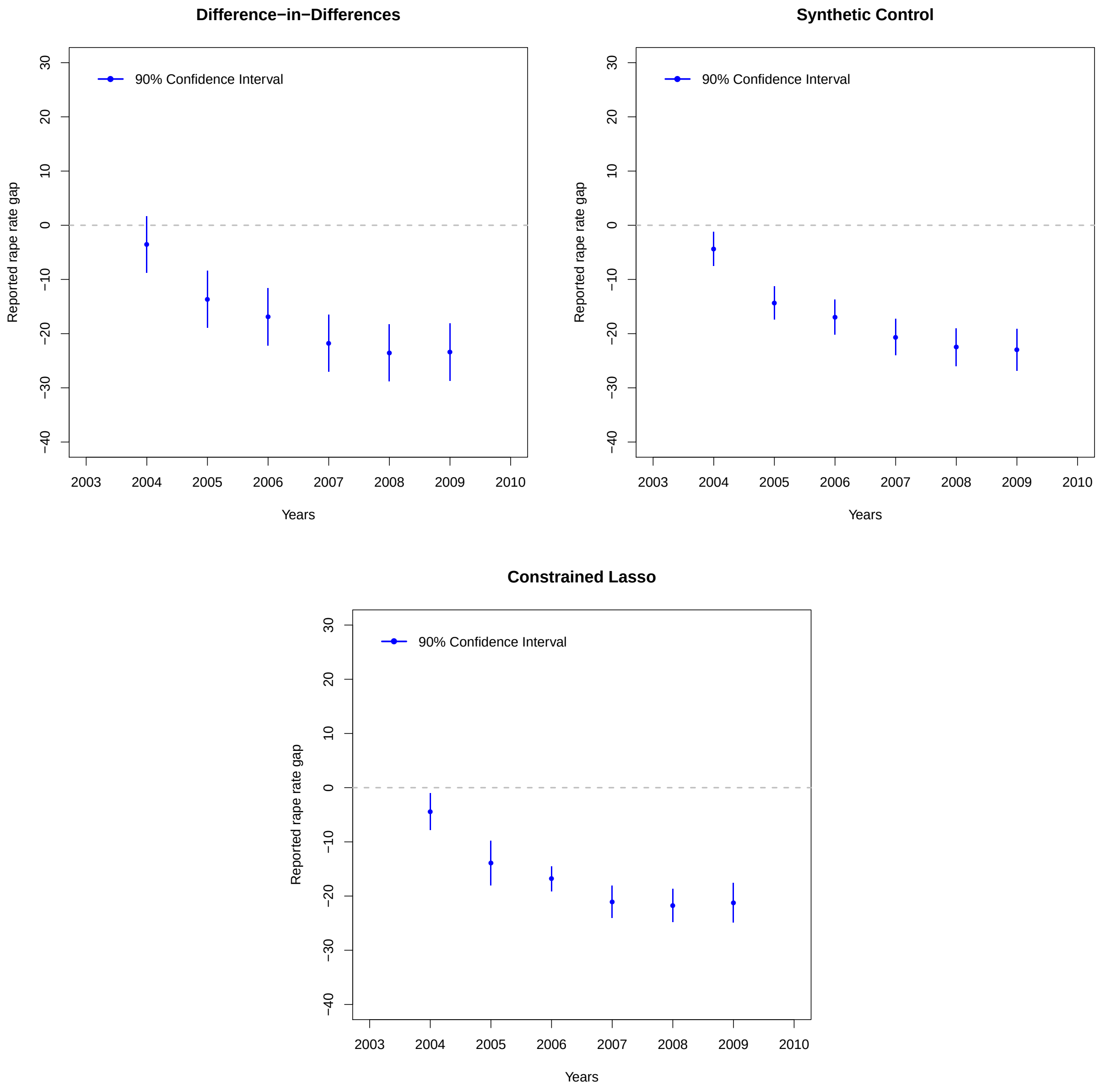


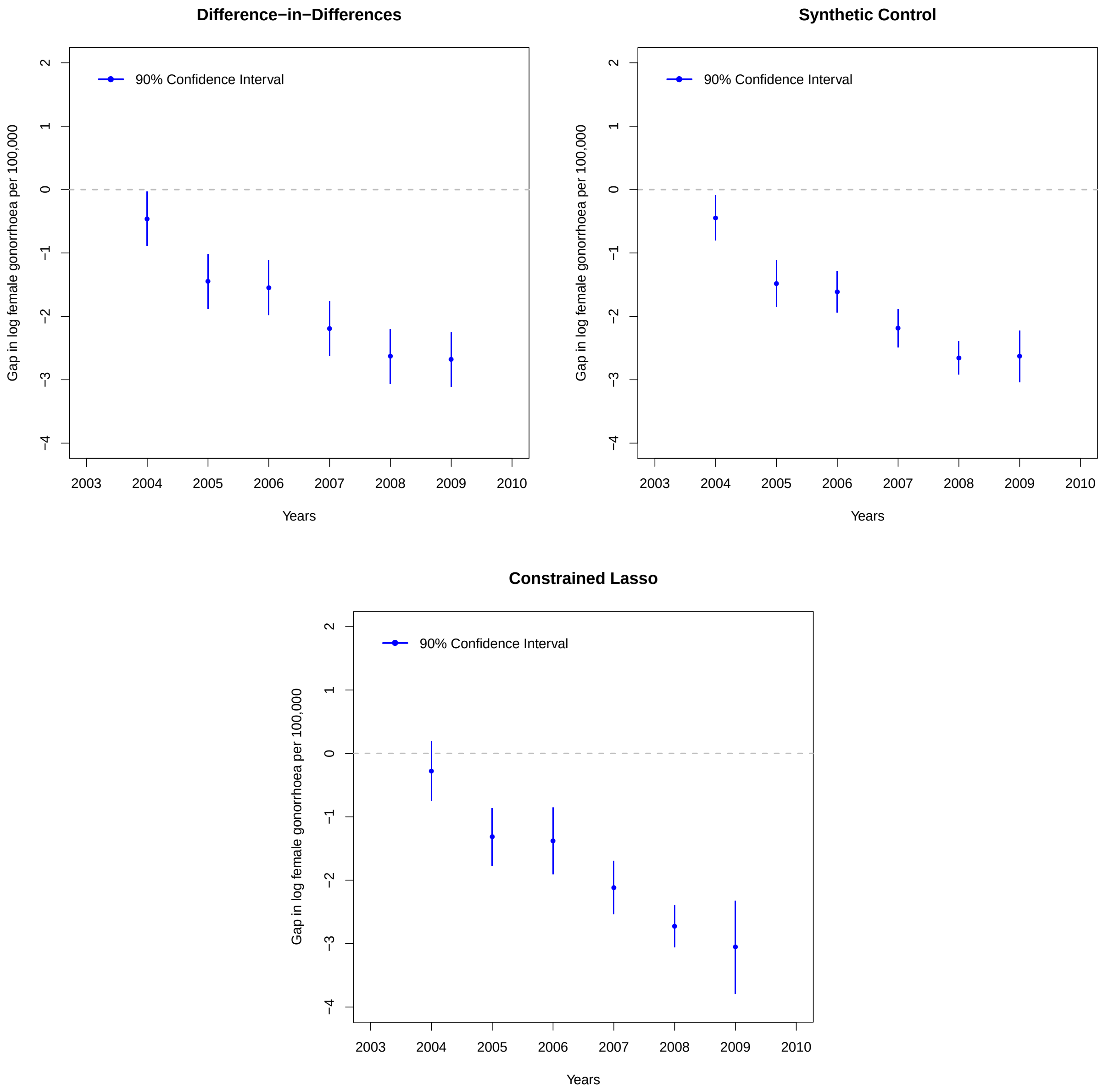
Figure 9: Pointwise Confidence Intervals Rape Rate


Difference−in−Differences

Synthetic Control

Constrained Lasso

Figure 10: Pointwise Confidence Intervals Gonorrhea


Difference−in−Differences
90% Confidence Interval
Gap in log female gonorrhoea per 100,000
Years
Synthetic Control
90% Confidence Interval
Gap in log female gonorrhoea per 100,000
Years
Constrained Lasso
90% Confidence Interval
Gap in log female gonorrhoea per 100,000
Years