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

Victor Chernozhukov*   Kaspar Wüthrich†   Yinchu Zhu‡

May 25, 2022

Abstract

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

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

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

†Department of Economics, University of California San Diego, 9500 Gilman Dr., La Jolla, CA 92093, USA; Email: kwuthrich@ucsd.edu

‡Brandeis University; 415 South Street, Waltham, MA 02453, USA; Email: yinchuzhu@brandeis.edu

1 Introduction

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

This paper provides generic and robust procedures for making inferences on policy effects estimated by CSC methods. We propose a general counterfactual modeling framework that nests and generalizes many traditional and new methods for counterfactual analysis. We focus on methods that are able to generate mean-unbiased proxies, P_t^N, for the counterfactual outcomes of the treated unit in the absence of the policy intervention, Y_{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 $\theta_t = Y_{1t}^I - Y_{1t}^N$, 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-treatment period, $\theta = \{\theta_t\}_{t=T_0+1}^{T_0+T_*}$. Specifically, we postulate a trajectory $\theta^0 = \{\theta_t^0\}_{t=T_0+1}^{T_0+T_*}$ and test the sharp null hypothesis that $\theta = \theta^0$. We also consider the problem of testing hypotheses about per-period effects θ_t and propose a simple algorithm for constructing pointwise confidence intervals via test inversion.

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

---

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

many classical and modern high-dimensional methods for estimating P_t^N.

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

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

(i) **Estimator Consistency and Stationary Weakly Dependent Errors**

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

(ii) **Estimator Stability and Stationary Weakly Dependent Data**

In practice, misspecification is an important concern. We show that even if the model for P_t^N is misspecified and the estimator of P_t^N, $\hat{P}_t^N$, is inconsistent, our procedures are still valid, provided that the data are stationary and weakly dependent and $\hat{P}_t^N$ satisfies a *stability* condition. This condition requires that $\hat{P}_t^N$ is stable under perturbations in a few observations. It is implied, for instance, if $\hat{P}_t^N$ is consistent for a pseudo-true parameter value but is shown to hold even in high-dimensional settings where consistency results under misspecification are often not available.

The main theoretical results in this paper are finite sample (non-asymptotic) bounds on the size accuracy of our methods; these bounds imply that our methods are exact as $T_0 \to \infty$. Unlike traditional asymptotic results, which are only informative when the sample

size is large enough, our non-asymptotic bounds show how different factors affect the finite sample performance. <u>This feature is relevant in CSC applications where sample sizes are often small.</u>

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

[Figure 1 around here.]

We make two additional contributions that may be of independent interest. First, we introduce the ℓ_1-constrained least squares estimator or *constrained Lasso* (e.g., Raskutti et al., 2011) as an essentially tuning-free alternative to existing penalized regression estimators and study its theoretical properties. Constrained Lasso nests SC and difference-in-differences, providing a unifying approach for the regression-based estimation of the mean proxies P_t^N. Second, we obtain theoretical consistency results for SC estimators in settings with potentially very many control units.

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

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

restrictions are correct but can cause power losses when they are not.

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

Related Literature. We contribute to the literature on inference procedures for CSC methods with few treated units. A popular method is the finite population permutation approach of Abadie et al. (2010), see also Firpo and Possebom (2018) and Abadie (2019). This approach permutes the policy assignment and relies on permutation distributions for inference. It corresponds to conventional randomization inference (Fisher, 1935) under random assignment of the policy (e.g., Abadie et al., 2010; Abadie, 2019). However, random assignment is not plausible in typical CSC applications, and assignment mechanisms are difficult to model and estimate when there are only few treated units (e.g., Abadie, 2019). Shaikh and Toulis (2019) propose randomization tests for settings with staggered treatment adoption, which encompass the approach of Abadie et al. (2010). The main assumption of Shaikh and Toulis (2019)'s approach is that policy adoption follows a Cox proportional hazards model. We do not model the assignment mechanism. Instead, we exploit stationarity and weak dependence of the errors across time in a repeated sampling framework. One advantage of exploiting the time series dimension is that we only require a suitable model for the potential outcome of the treated unit. By contrast, cross-sectional approaches often require estimating models for all units. That is, we only require a good "local" instead of a good "global" fit, which reduces the risk of model misspecification. On the other hand, our approach requires a large number of pre-treatment periods and relies on invariance of the error distribution under the intervention.

There is also an active literature on asymptotic inference methods for CSC models. Several papers focus on testing hypotheses about average or expected effects over time, requiring T_0 and T_* to be large. Li and Bell (2017), Carvalho et al. (2018), Chernozhukov et al. (2019), and Li (2020) introduce inference methods based on penalized and constrained regression methods. Arkhangelsky et al. (2018) propose inference methods for a version of

SC with time and unit weights, which admits a weighted regression formulation. Asymptotic inference methods based on factor and interactive FE models are proposed by Hsiao et al. (2012), Gobillon and Magnac (2016), Chan and Kwok (2016), Li and Bell (2017), Xu (2017), and Li (2018). Here we focus on sharp null hypotheses and permutation distributions and provide non-asymptotic performance guarantees. Our approach is generic and valid with many different methods, including constrained regression, factor models, and interactive FE estimators. Conley and Taber (2011) propose inference methods for difference-in-difference settings with few treated units. They exploit the cross-sectional dimension, relying on weak dependence and stationarity of the error terms across units, which may be hard to justify in typical CSC settings. By contrast, our procedures rely on stationarity and weak dependence of the errors over time. On the other hand, exploiting the time series dimension, our approach requires T_0 to be large, whereas Conley and Taber (2011) allow T_0 to be fixed. In related work, Cattaneo et al. (2021) provide prediction intervals for per-period effects estimated by SC methods. Their key observation is that there is both randomness from estimating the SC weights and from the prediction error. They propose a sampling-based inference method based on non-asymptotic probability bounds that accounts for both types of randomness and is valid with stationary and non-stationary data.

We recast the causal inference problem as a (counterfactual) prediction problem and build on the literature on conformal prediction (e.g., Vovk et al., 2005, 2009; Lei et al., 2013; Lei and Wasserman, 2014; Lei et al., 2018) and on the literature on permutation tests (e.g., Romano, 1990; Lehmann and Romano, 2005), which was started by Fisher (1935) in the context of randomization; see Rubin (1984) for a Bayesian justification. On a more general level, our approach is also connected to transformation-based approaches to model-free prediction (e.g., Politis, 2015). Let us discuss in more detail the relationship to Chernozhukov et al. (2018, CWZ18 henceforth), who extend classical conformal prediction to time series settings. Besides a different focus (prediction intervals for future outcome values vs. inference on policy effects), there are several important differences. First, we rely on permuting residuals, whereas CWZ18 permute blocks of data. Second, we theoretically analyze different types of permutations. In particular, we study the set of all permutations, which yields precise p-values in small samples. This set of permutations cannot be used in the frame-

work of CWZ18 unless the data are iid. Third, we allow for non-stationary data, whereas the prediction methods in CWZ18 are strictly limited to stationary data. Forth, CWZ18 rely on abstract high-level conditions on the test statistics and do not provide any primitive conditions. By contrast, we develop transparent sufficient conditions that can be verified for many traditional and modern CSC methods, and we provide explicit primitive conditions for a large selection of popular approaches. Finally, we establish the validity of our methods with time series data under misspecification and stability, whereas the theoretical results for weakly dependent data in CWZ18 require correct specification and consistency.

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

---

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

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

whereas Andrews (2003) focuses on low-dimensional GMM models.

Notation. For $q \geq 1$, the ℓ_q-norm of a vector is denoted by $\|\cdot\|_q$. We use $\|\cdot\|_0$ to denote the number of nonzero entries of a vector; $\|\cdot\|_\infty$ is used to denote the maximal absolute value of entries of a vector. We use the notation $a \lesssim b$ to denote $a \leq cb$ for some constant $c > 0$ that does not depend on the sample size. We use the notation $a \asymp b$ to denote $a \lesssim b$ and $b \lesssim a$. For a set A, $|A|$ denotes the cardinality of A. For any $a \in \mathbb{R}$, we define $\lfloor a \rfloor = \max\{z \in \mathbb{Z} : z \leq a\}$ and $\lceil a \rceil = \min\{z \in \mathbb{Z} : z \geq a\}$, where $\mathbb{Z}$ is the set of integers. We use $\mathbb{N}$ to denote the set of natural numbers.

2 A Conformal Inference Method

2.1 The Counterfactual Model

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

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

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

$\{Y_{1t}^I\} = \{Y_{1t}^N + \theta_t\}.^4$ *In other words, potential outcomes can be written as*

$$Y_{1t}^N = P_t^N + u_t \quad\bigg|\quad E(u_t) = 0, \quad t = 1, \ldots, T, \qquad\qquad \text{(CMF)}$$
$$Y_{1t}^I = P_t^N + \theta_t + u_t \quad\bigg|$$

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

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

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

Often, there is additional information in the form of untreated units, which can serve as controls. Specifically, suppose that there are $J \geq 1$ control units, indexed by $j = 2, \ldots, J+1$. 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, i.e., $Y_{jt} = Y_{jt}^N$ for $2 \leq j \leq J+1$ and $1 \leq t \leq T$. For each unit, we may also observe a vector of covariates X_{jt}. This motivates a variety of strategies for modeling and identifying P_t^N as discussed below.

---

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

[5]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.

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

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

$$H_0 : \ \theta = \theta^0, \tag{1}$$

where $\theta^0 = \left(\theta^0_{T_0+1}, \ldots, \theta^0_T\right)'$ 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 intervention in the post-treatment period since $Y^N_{1t} = Y^I_{1t} - \theta_t = Y_{1t} - \theta_t$. In the Appendix, we show that our method can also be used to test hypotheses about average effects.

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

$$Z_t = \begin{cases} \left(Y^N_{1t}, Y^N_{2t}, \ldots, Y^N_{J+1t}, X'_{1t}, \ldots, X'_{J+1t}\right)', & t \leq T_0 \\ \left(Y^I_{1t} - \theta^0_t, Y^N_{2t}, \ldots, Y^N_{J+1t}, X'_{1t}, \ldots, X'_{J+1t}\right)', & t > T_0. \end{cases}$$

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

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

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

Note that S is constructed such that high values indicate rejection. Different choices of

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

Remark 1 (Choice of Test Statistic)**.** While we focus on S_q, other test statistics can be used as well. For example, when capturing deviations in the average effect $T_*^{-1} \sum_{t=T_0+1}^{T} \theta_t$, it is useful to consider $S(\hat{u}) = T_*^{-1/2} \left| \sum_{t=T_0+1}^{T} \hat{u}_t \right|$. $\qquad\square$

We use (block) 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 permutations under study as Π and assume that Π contains the identity map $\mathbb{I}$. We focus on two different sets of permutations: (i) the set of all permutations, which we call *iid permutations*, Π_{all}, and (ii) the set of all (overlapping) *moving block permutations*, $\Pi_\rightarrow$.[6] The elements of $\Pi_\rightarrow$ are indexed by $j \in \{0, 1, \ldots, T-1\}$, and the permutation π_j is defined as

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

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

[Figure 2 around here.]

The choice of Π does not matter for the exact finite sample validity of our procedures if the residuals are exchangeable. However, Π_{all} has more elements than $\Pi_\rightarrow$, allowing us to compute more precise p-values and to test at lower significance levels. For the asymptotic validity under estimator consistency, the choice of Π depends on the assumptions that we are willing to impose on the stochastic shock sequence $\{u_t\}$ (cf. Section 3.1).

---

[6]We can also consider other types of permutations; for example, the "iid block" permutations. 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 "iid m-block" permutations Π_{mb}. In our context, choosing $m = T_*$ is natural, though other choices should work as well, similarly to the choice of block size in the time series bootstrap. We refer to CWZ18 for more results on block permutations.

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

Definition of p-Value. The p-value is

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

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

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

2.3 Models for Counterfactual Proxies via Synthetic Control and Panel Data

The availability of control units motivates several strategies for modeling the counterfactual mean proxies P_t^N. We estimate P_t^N based on the imputed data under the null hypothesis, $\mathbf{Z}(\theta^0)$, and write Y_{1t}^N instead of $Y_{1t}^I - \theta_t^0$ to alleviate the exposition.

2.3.1 Difference-in-Differences Methods

The difference-in-differences method postulates the following model for the counterfactual mean proxy (e.g., Doudchenko and Imbens, 2016, Section 5.1): $P_t^N = \mu + \frac{1}{J}\sum_{j=2}^{J+1} Y_{jt}^N$. This model automatically embeds the identifying information. The counterfactual mean proxy can be estimated as $\hat{P}_t^N = \frac{1}{T}\sum_{s=1}^{T}\left(Y_{1s}^N - \frac{1}{J}\sum_{j=2}^{J+1} Y_{js}^N\right) + \frac{1}{J}\sum_{j=2}^{J+1} Y_{jt}^N$.

---

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

2.3.2 Synthetic Control and Constrained Lasso

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

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

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

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

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

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

As an alternative, we can consider the more flexible model[10]

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

---

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

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

[10]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 the formal properties of these estimators. Here we emphasize another version of relaxing SC, model (5), which leads to constrained Lasso (6). Constrained Lasso demonstrates an excellent theoretical and practical performance: it is tuning-free, performs very well empirically and in simulations, and we prove that it is consistent for dependent data without any sparsity conditions on the weights and that it satisfies the estimator stability condition required for validity under misspecification. We emphasize that this estimator generally differs from the cross-validated Lasso estimator.

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

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

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

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

2.3.3   Penalized Regression Methods

Consider a linear model for P_t^N: $P_t^N = \mu + \sum_{j=2}^{J+1} w_j Y_{jt}^N$. We maintain the identifying assumption (SC). The counterfactual is estimated by $\hat{P}_t^N = \hat{\mu} + \sum_{j=2}^{J+1} \hat{w}_j Y_{jt}^N$, where

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

and $\mathcal{P}(w)$ is a penalty function that penalizes deviations away from zero. If it is desired to penalize deviations away from other focal points w^0, for example, $w^0 = (1/J, \ldots, 1/J)$ used in the difference-in-differences approach, we may always use instead: $\mathcal{P}(w) \leftarrow \mathcal{P}(w - w^0)$. Note that it is straightforward to incorporate covariates X_{jt} into the estimation problem (7).

Different variants of $\mathcal{P}(w)$ can be considered. Examples include: Lasso (Tibshirani, 1996), where $\mathcal{P}(w) = \lambda \|w\|_1$ and λ is a tuning parameter; Elastic Net (Zou and Hastie,

2005), where $\mathcal{P}(w) = \lambda\left((1-\alpha)\|w\|_2^2 + \alpha\|w\|_1\right)$ and λ and α are tuning parameters; Lava (Chernozhukov et al., 2017), where $\mathcal{P}(w) = \inf_{a+b=w} \lambda\left((1-\alpha)\|a\|_2^2 + \alpha\|b\|_1\right)$ and λ and α are tuning parameters.

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

2.3.4 Interactive Fixed Effects, Factor, and Matrix Completion Models

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

$$Y_{jt}^N = \lambda_j' F_t + X_{jt}'\beta + u_{jt}, \quad \text{for} \quad 1 \le j \le J+1 \text{ and } 1 \le t \le T, \tag{8}$$

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

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

The model leads to the following proxy:

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

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

Hsiao et al. (2012) appears to the be first work that proposed the use of factor models for predicting the (missing) counterfactual responses specifically in SC settings. Gobillon and Magnac (2016) and Xu (2017) employ Bai (2009)'s estimator in this setting, albeit provide no

formally justified inference methods. Formal inference results for interactive FE and factor models in SC designs are developed in Chan and Kwok (2016) and Li (2018) among others.[11]

Other recent applications to predicting counterfactual responses include Amjad et al. (2018) and Athey et al. (2018) (using, respectively, singular value thresholding and the nuclear norm penalization).[12] Our method delivers a way to perform valid inference for policy effects using any of the factor model estimators used in these proposals applied to the complete data under the null.[13] We shall be focusing on Bai (2009)'s alternating least squares estimator[14] and on matrix completion via nuclear norm penalization when verifying our conditions.

2.4 Models for Counterfactual Proxies via Time Series and Fused Models

2.4.1 Simple Time Series Models

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

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

---

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

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

[13]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 FE model to estimate the model for the entire data under the null hypothesis.

[14]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.

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

In model (10), the mean unbiased proxy is given by $P_t^N = \mu + \rho(Y_{1(t-1)}^N - \mu)$. Note that the policy effect here is transitory, namely it does not feed-forward itself on the future values of Y_{1t}^I beyond the current values.[16] Under the null hypothesis, we can impute the unobserved counterfactual as $Y_{1t}^N = Y_{1t} - \theta_t$ and estimate the model using traditional time series methods, and we can conduct inference by permuting the residuals.

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

2.4.2 Fused Time-Series/Panel Models

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

$$
\begin{array}{c|c|c}
\begin{aligned}
Y_{1t}^N &= C_t^N + \varepsilon_t \\
Y_{1t}^I &= C_t^N + \theta_t + \varepsilon_t
\end{aligned}
&
\begin{aligned}
\varepsilon_t &= \rho(\varepsilon_{t-1}) + u_t, \ \{u_t\} \text{ iid}, \ E(u_t) = 0, \\
\{u_t\} &\text{ is independent of } \{C_t^N\},
\end{aligned}
& t = 1, \ldots, T,
\end{array}
\qquad (11)
$$

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

---

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

3 Theory

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

3.1 Approximate Validity under Estimator Consistency

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

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

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

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

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

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

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

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

Assumption 3 imposes weak and easy-to-verify conditions on the performance of the estimators $\hat{P}_t^N$ of the counterfactual mean proxies P_t^N. These conditions are readily implied by the existing results for many estimators discussed in Section 2. In Section 4, we provide explicit primitive conditions and references to primitive conditions implying Assumption 3.[17]

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

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

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

The above bound is non-asymptotic, allowing us to claim uniform validity with respect to a rich variety of data generating processes. Using simulations and empirical examples, we

---

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

verify that our tests have good power and generate meaningful empirical results. There are other considerations that also affect power. For example, the better the model for P_t^N, the less variance the stochastic shocks will have, subject to assumed invariance to the policy. The smaller the variance of the shocks, the more powerful the testing procedure will be.

3.2   Approximate Validity under Estimator Stability

Misspecification is an important practical concern, and consistency of the estimators of the counterfactual mean proxies P_t^N may be questionable in certain settings. The classical analysis of misspecification focuses on convergence to pseudo-true values (e.g., White, 1996). If it is possible to show that the estimator of the counterfactual mean proxy, $\hat{P}_t^N$, is consistent for some pseudo-true value P_t^{N*} and that $\left\{Y_{1t}^N - P_t^{N*}\right\}_{t=1}^T$ is stationary and weakly dependent, the theoretical results in Section 3.1 imply the validity of our procedure. Pseudo-true consistency can often be verified for low-dimensional models, but consistency results under misspecification remain elusive in high-dimensional settings. Therefore, we consider a notion of approximate exchangeability, which only requires the estimator to be *stable* instead of consistent for a pseudo-true value. This stability condition does not require $\hat{P}_t^N$ to be consistent for anything, nor does it rely on correct specification of the counterfactual mean proxies. In the Appendix, we illustrate the difference between consistency and stability based on the analytically tractable example of Ridge regression.

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

We now formalize the notion of stability of an estimator. To emphasize the dependence of $S(\hat{u})$ on the estimator, with a slight abuse of notation, we write $S(\mathbf{Z}, \beta) = \phi(Z_{T_0+1}, \ldots, Z_{T_0+T_*}; \beta)$. Let $\{\tilde{Z}_t\}_{t=1}^T$ be iid from the distribution of Z_1 and independent of $\mathbf{Z}$. For any $H \subset \{1, \ldots, T\}$, 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 $\{\tilde{Z}_t\}_{t \in H}$.

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

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

Assumption 4 (Estimator Stability). *Let $\Pi = \Pi_\rightarrow$. There exist non-decreasing functions $\varrho_T(\cdot)$ such that $P\left(\max_{\pi \in \Pi}\left|S\left(\mathbf{Z}^\pi, \hat{\beta}(\mathbf{Z})\right) - S\left(\mathbf{Z}^\pi, \hat{\beta}(\mathbf{Z}_H)\right)\right| \leq \varrho_T(|H|)\right) \geq 1 - \gamma_{1,T}$ and $P\left(\max_{\pi \in \Pi}\left|S\left((\dot{\mathbf{Z}})^\pi, \hat{\beta}(\mathbf{Z})\right) - S\left((\dot{\mathbf{Z}})^\pi, \hat{\beta}(\mathbf{Z}_H)\right)\right| \leq \varrho_T(|H|)\right) \geq 1 - \gamma_{1,T}$ for any $H \in \{\widetilde{H}_1, \ldots, \widetilde{H}_R\}$, where $\dot{\mathbf{Z}} \stackrel{d}{=} \mathbf{Z}$ and $\dot{\mathbf{Z}}$ is independent of $(\mathbf{Z}, \{\tilde{Z}_t\}_{t=1}^T)$.*

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

Let $\Psi(x; \beta) = P(\phi(Z_{T_0+1}, \ldots, Z_{T_0+T_*}; \beta) \leq x)$. Our strategy is to show that, under the null hypothesis, $\hat{F}(\phi(Z_{T_0+1}, \ldots, Z_{T_0+T_*}; \hat{\beta}(\mathbf{Z})))$ is approximately uniform on $(0, 1)$. We exploit the stability condition in Assumption 4 and show that $\hat{F}(\phi(Z_{T_0+1}, \ldots, Z_{T_0+T_*}; \hat{\beta}(\mathbf{Z})))$ can be approximated by $\Psi\left(\phi(\bar{Z}_{T_0+1}, \ldots, \bar{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)$, which has the uniform

distribution on (0,1). Here $(\bar{Z}_{T_0+1}, \ldots, \bar{Z}_{T_0+T_*})$ 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 $(Z_{T_0+1}, \ldots, Z_{T_0+T_*})$.

We impose the following regularity conditions on the data.

Assumption 5 (Regularity of the Data). *The data under the null, $\{Z_t\}_{t=1}^T$, are stationary and β-mixing with coefficient $\beta_{\mathrm{mixing}}(\cdot)$ satisfying $\beta_{\mathrm{mixing}}(i) \leq D_1 \exp(-D_2 i^{D_3})$ for some constants $D_1, D_2, D_3 > 0$. For $1 \leq j \leq R$, there exist sequences $\xi_T > 0$ and $\gamma_{2,T} = o(1)$ such that $P\left(\sup_{x \in \mathbb{R}} \left| \partial \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) / \partial x \right| \leq \xi_T\right) \geq 1 - \gamma_{2,T}$.*

Stationarity and β-mixing are commonly imposed conditions on time series data. For a large class of Markov chains, GARCH and various stochastic volatility models, $D_3 = 1$ (cf. Carrasco and Chen, 2002). Let $(\dot{Z}_{T_0+1}, \ldots, \dot{Z}_{T_0+T_*})$ 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 density 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 satisfying the SC restrictions, the density of $Y_t - X_t'w$ is bounded by a constant that does not depend on w.

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

Theorem 2 (Approximate Validity under Estimator Stability). *Let $\Pi = \Pi_\rightarrow$. Suppose that Assumptions 4 and 5 hold. Then, under the null hypothesis, there exists a constant $C_1 > 0$ depending only on D_1, D_2 and D_3 such that for any R with $k < \lfloor T_0/R \rfloor$ and $R < T_0/2$,*

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

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 $(0, 1)$; with iid or

exchangeable data, $\hat{p}$ conditional on $\hat{\beta}(\mathbf{Z})$ has an exact uniform distribution. Therefore, we can view Theorem 2 as a result for approximate exchangeability.

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

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

4 Sufficient Conditions for Consistent Estimation

In this section, we revisit the representative models of counterfactual proxies introduced in Section 2. Primitive conditions are provided to guarantee that the estimation of the counterfactual mean proxies is accurate enough for the asymptotic validity of the proposed procedure. In particular, these conditions can be used to verify Assumption 3. The regularity conditions (e.g., bounded moments, weak serial dependence) for different models are stated in the Appendix and are commonly imposed in the literature for these models. The counterfactual mean proxies P_t^N are estimated based on the imputed data under the null, $\mathbf{Z}(\theta^0)$, and we write Y_{1t}^N instead of $Y_{1t}^I - \theta_t^0$ to alleviate the exposition. All the results in this section assume that $T_0 \to \infty$ and $J \to \infty$ (if J is present in the model).

4.1 Difference-in-Differences

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

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

4.2   Synthetic Control and Constrained Lasso

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

To provide a unified framework for these models, we use Y to denote a generic vector of outcomes and X to denote the design matrix throughout this section. For example, in Section 2, we set $Y = Y_1^N$ and $X = (Y_2^N, \ldots, Y_{J+1}^N)$, 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

$$Y = Xw + u, \tag{12}$$

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

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

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

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

The constrained Lasso estimator is

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

23

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

In general, we choose the parameter space $\mathcal{W}$ to be an arbitrary subset of an ℓ_1-ball with bounded radius. The following result gives very mild conditions under which the constrained least squares estimators are consistent and satisfy Assumption 3.[18]

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

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

4.3   Models with Factor Structures

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

---

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

4.3.1 Pure Factor Models

Recall from Section 2.3.4 the standard factor model $Y_{jt}^N = \lambda_j' F_t + u_{jt}$, where $F = (F_1, \ldots, F_T)' \in \mathbb{R}^{T \times k}$ 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 $P_t^N = \lambda_1' F_t$. We identify P_t^N by imposing the condition that the idiosyncratic terms and the factor structure are uncorrelated (cf. Condition (FE)).

We use the standard principal component analysis (PCA) for estimating P_t^N.[19] Let $Y^N \in \mathbb{R}^{T \times N}$ be the matrix whose (t, j) entry is Y_{jt}^N. We compute $\hat{F} = (\hat{F}_1, \ldots, \hat{F}_T)' \in \mathbb{R}^{T \times k}$ 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 2 (Pure Factor Model). *Assume standard regularity conditions given in Bai (2003), including the identification condition (FE). Consider the factor model and the principal component estimator. Then, for any $1 \leq t \leq T$, as $N \to \infty$ and $T \to \infty$, we have $\hat{P}_t^N - P_t^N = O_P(1/\sqrt{N} + 1/\sqrt{T})$ and $\frac{1}{T} \sum_{t=1}^T (\hat{P}_t^N - P_t^N)^2 = O_P(1/N + 1/T)$.*

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

4.3.2 Factor plus Regression Model: Interactive FE Model

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

---

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

the condition that the idiosyncratic terms are independent of the factor structure and the observed covariates (cf. Condition (FE)).

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

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

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

Lemma 3 (Interactive FE Model). *Assume the standard conditions in Bai (2009), including the identification condition (FE). Then, for any $1 \leq t \leq T$, $\hat{P}_t^N - P_t^N = O_P(1/\sqrt{T} + 1/\sqrt{N})$ and $\frac{1}{T} \sum_{t=1}^{T} (\hat{P}_t^N - P_t^N)^2 = O_P(1/T + 1/N)$.*

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

4.3.3 Matrix Completion via Nuclear Norm Regularization

Suppose that

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

where M_{jt} is the (j, t)-element of an unknown matrix $M \in \mathbb{R}^{(J+1) \times T}$ satisfying $\|M\|_* \leq K$, where $\| \cdot \|_*$ denotes the nuclear norm (the sum of singular values). We observe Y_{jt}^N for $(j, t) \in \{1, \ldots, T\} \times \{1, \ldots, J+1\} \backslash \{(1, t) : T_0 + 1 \leq t \leq T\}$. The identifying condition is that $E(u \mid M) = 0$ and that conditional on M, $\{u_j\}_{j=1}^{J+1}$ 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 on matrix completion considers the model (15) under the assumption of missingness at random and exploits the assumption that the rank of M is

low.[20] Recently, Athey et al. (2018) introduce this method to study treatment effects in panel data models and point out the unobserved counterfactuals correspond to entries that are missing in a very special pattern, rather than at random. Assuming the usual low rank condition on M, they employ the nuclear norm penalized estimator and provide bounds on the estimation error in the typical setup of causal panel data models.

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

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

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

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

The result is notable because no sub-Gaussian assumptions are required. The estimator in (16) does not explicitly require a low-rank condition on M. Instead, we impose a growth restriction on K. When M is generated by a strong factor structure and the null hypothesis

---

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

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

contains 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 $c > 0$. In the case of weak factors, we can choose $K \ll \sqrt{NT}$ and obtain consistency.

4.4 Time Series and Fused Models

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

4.4.1 Autoregressive Models

The linear autoregressive model with K lags can be written as $Y_{1t}^N = \rho_0 + \sum_{j=1}^K \rho_j Y_{1t-j}^N + u_t$, where $\{u_t\}_{t=1}^T$ is an iid sequence with $E(u_t) = 0$.[22] The counterfactual proxy for Y_{1t}^N is $P_t^N = \rho_0 + \sum_{j=1}^K \rho_j Y_{1t-j}^N$. We write P_t^N as $P_t^N = y_t' \rho$, where $y_t = (1, Y_{1t-1}^N, Y_{1t-2}^N, \ldots, Y_{1t-K}^N)' \in \mathbb{R}^{K+1}$ and $\rho = (\rho_0, \ldots, \rho_K)' \in \mathbb{R}^{K+1}$. 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 estimator for P_t^N is $\hat{P}_t^N = y_t' \hat{\rho}$.

Lemma 5 (Linear AR Model). *Suppose that $\{u_t\}_{t=1}^T$ is an iid sequence with $E(u_1) = 0$ and $E(u_1^4)$ uniformly bounded and the roots of $1 - \sum_{j=1}^K \rho_j L^j = 0$ are uniformly bounded away from the unit circle. Then, for any $T_0 + 1 \le t \le T$, $\hat{P}_t^N - P_t^N = o_P(1)$ and $\frac{1}{T} \sum_{t=K+1}^T (\hat{P}_t^N - P_t^N)^2 = o_P(1)$.*

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

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

Lemma 6 (Nonlinear AR Model). *Suppose that (1) $\|\hat{\rho} - \rho\| = O_P(r_T)$ with $r_T = o(1)$ for some appropriate norm $\|\cdot\|$ and $\max_{K+1 \le t \le T} |\hat{\rho}(Y_{1t-1}^N, Y_{1t-2}^N, \ldots, Y_{1t-K}^N) - \rho(Y_{1t-1}^N, Y_{1t-2}^N, \ldots, Y_{1t-K}^N)| \le \ell_T \|\hat{\rho} - \rho\|$ for some $\ell_T r_T = o(1)$. Then, for any $T_0 + 1 \le t \le T$, $\hat{P}_t^N - P_t^N = o_P(1)$ and $\frac{1}{T} \sum_{t=K+1}^T (\hat{P}_t^N - P_t^N)^2 = o_P(1)$.*

---

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

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

4.4.2 Fused Panel/Time Series Models with AR Errors

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

We estimate P_t^N via a two-stage procedure. In the first stage, we estimate C_t^N using the techniques we considered before and obtain say $\hat{C}_t^N$. In the second stage, we estimate $\rho(\varepsilon_{t-1})$ by fitting an autoregressive model to the estimated residuals $\{\hat{\varepsilon}_t\}_{t=1}^T$, where $\hat{\varepsilon}_t = Y_{1t}^N - \hat{C}_t^N$. For simplicity, we consider a linear model in the second stage estimation. Analogous results can be obtained for more general models. To be specific, assume that $\varepsilon_t = x_t'\rho + u_t$, where $x_t = (\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-K})' \in \mathbb{R}^K$ 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 7 (AR Errors). *Suppose that $\{u_t\}_{t=1}^T$ is an iid sequence with $E(u_t) = 0$ and $E(u_1^4)$ uniformly bounded and the roots of $1 - \sum_{j=1}^K \rho_j L^j = 0$ are uniformly bounded away from the unit circle. We assume that (1) $\sum_{t=1}^T (\hat{C}_t^N - C_t^N)^2 = o_P(T)$, and (2) $\hat{C}_t^N - C_t^N = o_P(1)$ for $T_0 - K + 1 \le t \le T$. Then, for any $T_0 + 1 \le t \le T$, $\hat{P}_t^N - P_t^N = o_P(1)$ and $\sum_{t=K+1}^T \left(\hat{P}_t^N - P_t^N\right)^2 = o_P(T)$*

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

5 Empirical Application

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

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

[Figure 3 around here.]

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

Before turning to the main results, we use the placebo tests proposed in the Appendix to assess the plausibility of the underlying assumptions. Specifically, based on the pre-treatment data, we test $H_0 : \theta_{2003-\tau+1} = \cdots = \theta_{2003} = 0$ for $\tau \in \{1, 2, 3\}$. Rejections of this null undermine the credibility of the assumptions underlying our procedure and the inferences on policy effects in the post-treatment period. Table 1 presents the results. Figure 4 complements the formal tests with plots of the residuals from fitting the three models to the pre-treatment data. The placebo tests and the residual plots provide evidence in favor of the credibility of our inference method in conjunction with SC and, especially, constrained Lasso, but suggest that the difference-in-differences results need to be interpreted with caution.

---

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

[Table 1 around here.]

[Figure 4 around here.]

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

$$H_0 : \theta_{2004} = \theta_{2005} = \cdots = \theta_{2009} = 0. \tag{17}$$

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

[Table 2 around here.]

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

[Figure 5 around here.]

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

[Figure 6 around here.]

Acknowledgements

References

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

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

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

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

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

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

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

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

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

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

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

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

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

Berbee, H. (1987). Convergence rates in the strong law for bounded mixing sequences. *Probability Theory and Related Fields*, 74(2):255–270.

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):717–772.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, 23(2):169–192.

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

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

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*, 113(523):1094–1111.

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

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

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

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

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

McCarthy, C. A. (1967). C_p. *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.

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

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

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

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.

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

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

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

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

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

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

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

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.

Ye, F. and Zhang, C.-H. (2010). Rate minimaxity of the lasso and dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research*, 11(Dec):3519–3540.

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

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

Figures Main Text

Figure 1: Small Sample Size Properties (Nominal Level: 10%)
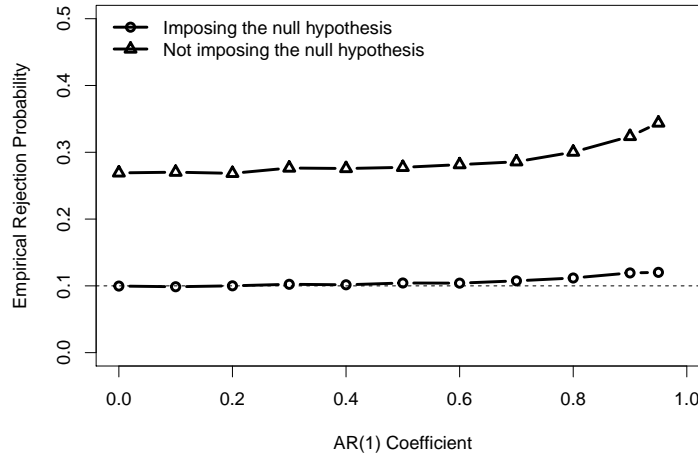
Notes: Empirical rejection probability from testing $H_0 : \theta_{T_0+1} = 0$. The data are generated as $Y_{1t}^N = \sum_{j=2}^{J+1} w_j Y_{jt}^N + u_t$, where $Y_{jt}^N \sim N(0,1)$ is iid across (j,t), $\{u_t\}$ is a Gaussian AR(1) process, $(w_2, \ldots, w_{J+1})' = (1/3, 1/3, 1/3, 0, \ldots, 0)'$, $T_0 = 19$, and $J = 50$. The weights are estimated using the canonical SC method (cf. Section 2.3.2).

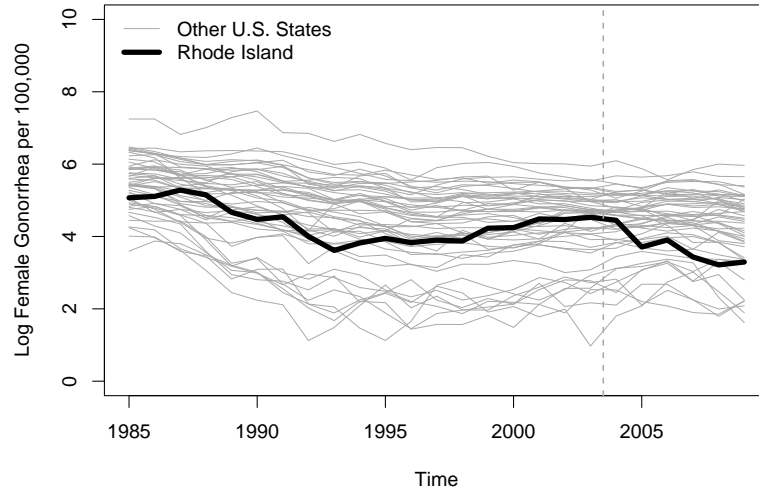Figure 2: Graphical Illustration Permutations
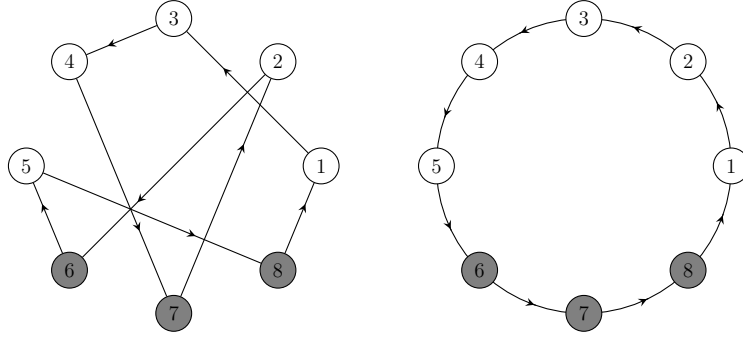
Notes: The left figure gives an example of an iid permutation of $\{1, 2, 3, 4, 5, 6, 7, 8\}$. The right figure gives an example of a moving block permutation of $\{1, 2, 3, 4, 5, 6, 7, 8\}$. $T_0 = 5$, $T_* = 3$. Pre-treatment periods are white; post-treatment periods are gray.

Figure 3: Raw Data

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

Figure 4: Graphical Placebo Checks
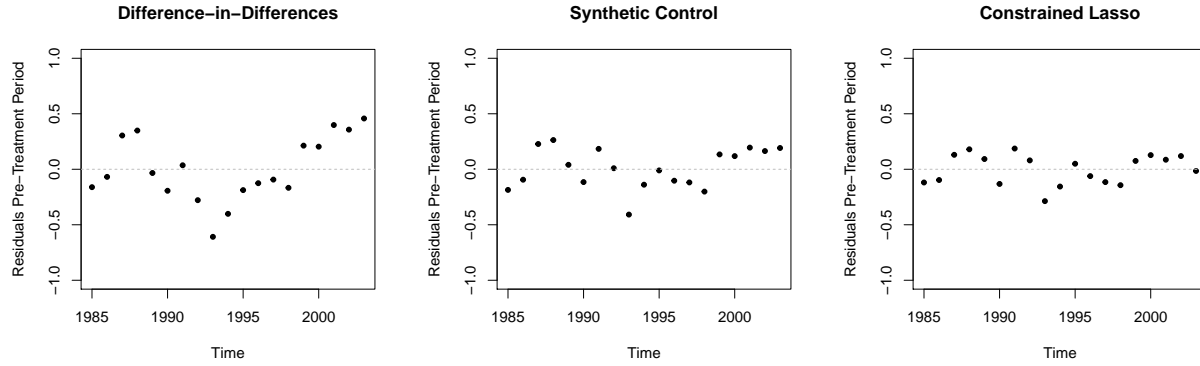
Notes: Data are from Cunningham and Shah (2018). The figure plots the pre-treatment residuals estimated using difference-in-differences, SC, and constrained Lasso.

Figure 5: Pointwise Confidence Intervals
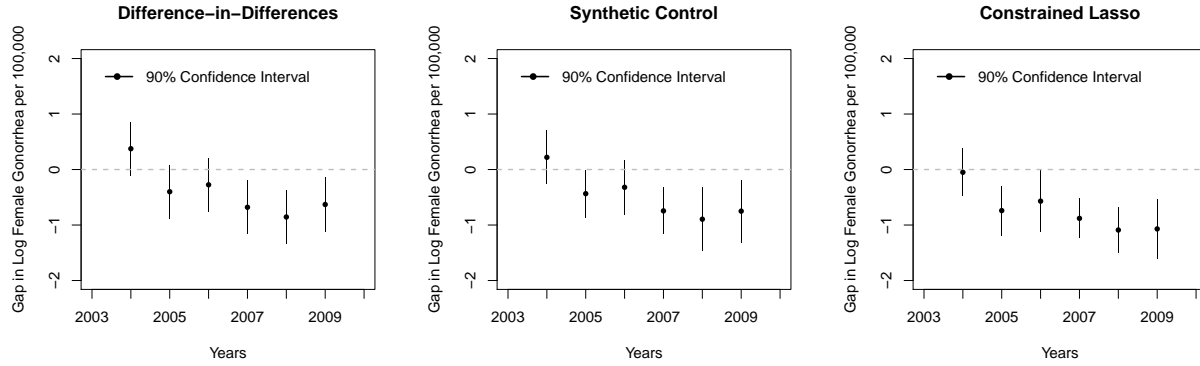
Notes: Data are from Cunningham and Shah (2018). The figure plots pointwise 90% confidence intervals computed using Algorithm 1.

Figure 6: Leave-one-out Robustness Checks
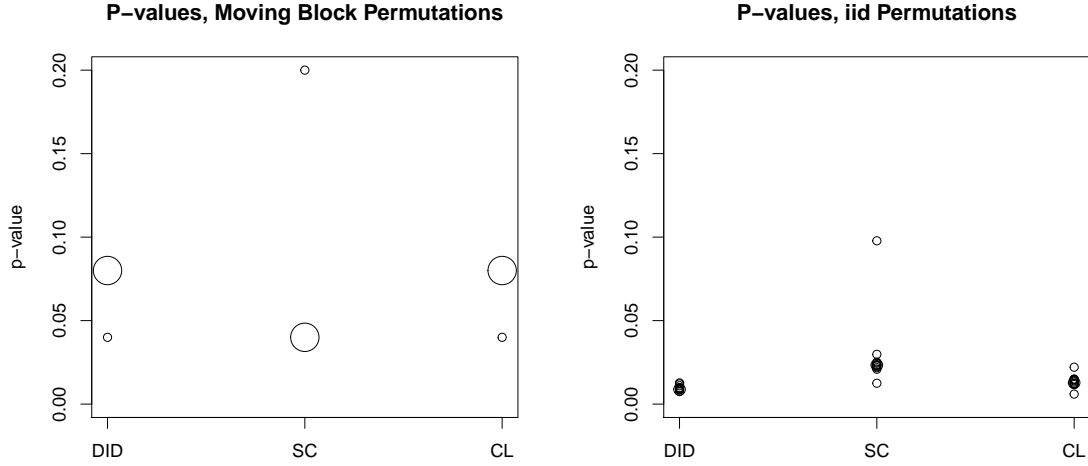
Notes: Data are from Cunningham and Shah (2018). This figure shows the distribution of p-values from testing null hypothesis (17), leaving-out one of the control states with non-zero weight at the time. The size of the circles is proportional to the number of p-values. DID: difference-in-differences; SC: synthetic control; CL: constrained Lasso.

Tables Main Text

Table 1: Placebo Specification Tests

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

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

Table 2: Zero Effect Null Hypothesis

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

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

Victor Chernozhukov, Massachusetts Institute of Technology. Email: vchern@mit.edu
Kaspar Wüthrich, University of California San Diego. Email: kwuthrich@ucsd.edu
Yinchu Zhu, Brandeis University. Email: yinchuzhu@brandeis.edu

A Extensions

A.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, $\bar{\theta} = T_*^{-1} \sum_{t=T_0+1}^{T} \theta_t$:

$$H_0 : \bar{\theta} = \bar{\theta}^0 \tag{A.1}$$

Hypothesis (A.1) can be tested by collapsing the data into averages of non-overlapping blocks of T_* observations over the time dimension. To simplify the exposition, we assume that T/T_* is an integer. 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 $\bar{Y}_{1r}^I = T_*^{-1} \sum_{t=r}^{r+T_*-1} Y_t^I$:

$$\left. \begin{array}{l} \bar{Y}_{1r}^N = \bar{P}_r^N + \bar{u}_r \\ \bar{Y}_{1r}^I = \bar{P}_r^N + \bar{\theta}_r + \bar{u}_r \end{array} \right| \quad E(\bar{u}_r) = 0, \quad r = 1, T_* + 1, 2T_* + 1, \ldots, T_0 - T_* + 1, T_0 + 1,$$

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 (collapsed) data under the null as $\bar{\mathbf{Z}} = (\bar{Z}_1, \ldots, \bar{Z}_{T_0+1})'$, where

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

and $\bar{X}_{jr} = T_*^{-1} \sum_{t=r}^{r+T_*-1} X_{jt}$ for $j = 1, \ldots, J+1$. Note that testing hypothesis (A.1) is equivalent to testing a hypothesis concerning a per-period effect based on the aggregated data $\bar{\mathbf{Z}}$. Specifically, we estimate the average proxy $\hat{\bar{P}}_r^N$ based on the aggregated data $\bar{\mathbf{Z}}$ and obtain the residuals $\hat{\bar{u}} = (\hat{\bar{u}}_1, \hat{\bar{u}}_{T_*+1}, \ldots, \hat{\bar{u}}_{T_0+1})'$, where $\hat{\bar{u}}_r = \bar{Y}_{1r}^N - \hat{\bar{P}}_r^N$. The test statistic is $S\left(\hat{\bar{u}}\right)$, and p-values can be obtained based on permutations of $\hat{\bar{u}}$ as described in Section 2.2.

The key assumption underlying this procedure is that the average mean proxy $\bar{P}_r^N$ can be identified and consistently estimated based on the aggregated data $\bar{\mathbf{Z}}$. This is the case for SC and the other regression-based estimators discussed in Sections 2.3.2–2.3.3, provided that $E(\bar{u}_r \bar{Y}_{jr}^N) = 0$ for $2 \leq j \leq J+1$ and that the sufficient conditions for consistent estimation in Section 4.2 hold for the aggregate data. By contrast, identification and estimation of $\bar{P}_r^N$ may not be possible for nonlinear and dynamic models. Under consistent estimation of $\hat{\bar{P}}_r^N$, the formal properties of the test follow from the results in Section 3.1 since stationarity and weak dependence of $\{u_t\}$ imply stationarity and weak dependence of $\{\bar{u}_r\}$. Alternatively, if $\hat{\bar{P}}_r^N$ can be shown to be stable and the aggregate data are stationary and weakly dependent, the properties of the test follow from the results in Section 3.2. Finally, we emphasize that the effective sample size is T/T_* instead of T, such that T needs to be substantially larger than T_*.

A.2 Multiple Treated Units

Our method can be extended to accommodate multiple treated units by collapsing the data into averages across the treated units. Consider a setup with L treated units, indexed by $j = 1, \ldots, L$, and J control units, indexed by $j = L+1, \ldots, J+L$. Suppose that Assumption 1 holds for all treated units:

$$\begin{aligned} Y_{jt}^N &= P_{jt}^N + u_{jt} \\ Y_{jt}^I &= P_{jt}^N + \theta_{jt} + u_{jt} \end{aligned} \quad \Bigg| \quad E(u_{jt}) = 0, \quad t = 1, \ldots, T, \quad j = 1, \ldots, L.$$

Under this assumption, hypotheses about the unit-specific policy effects $\{\theta_{jt}\}$ can be tested by separately applying the proposed inference procedure to each treated unit. In addition, one is often also interested in conducting inferences about the average treatment effects on

the treated units, $\{\bar{\theta}_t\}$, where $\bar{\theta}_t = L^{-1} \sum_{j=1}^{L} \theta_{jt}$.

Specifically, consider the following null hypothesis:

$$H_0 : \left(\bar{\theta}_{T_0+1}, \ldots, \bar{\theta}_T\right) = \left(\bar{\theta}^0_{T_0+1}, \ldots, \bar{\theta}^0_T\right). \tag{A.2}$$

To test hypothesis (A.2), 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 $\bar{Y}_t^I = L^{-1} \sum_{j=1}^{L} Y_{jt}^I$:

$$\begin{array}{c|c} \bar{Y}_t^N = \bar{P}_t^N + \bar{u}_t \\ \bar{Y}_t^I = \bar{P}_t^N + \bar{\theta}_t + \bar{u}_t \end{array} \quad E(\bar{u}_t) = 0, \quad t = 1, \ldots, T,$$

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 $\bar{\mathbf{Z}} = (\bar{Z}_1, \ldots, \bar{Z}_T)'$, where

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

and $\bar{X}_t = L^{-1} \sum_{j=1}^{L} X_{jt}$. To test hypothesis (A.2), we compute the estimated average proxy $\hat{\bar{P}}_t^N$ based on the aggregated data $\bar{\mathbf{Z}}$ and obtain the residuals $\hat{\bar{u}} = (\hat{\bar{u}}_1, \ldots, \hat{\bar{u}}_T)'$, where $\hat{\bar{u}}_t = \bar{Y}_t^N - \hat{\bar{P}}_t^N$ for $t = 1, \ldots, T$. The test statistic is $S\left(\hat{\bar{u}}\right)$, and p-values can be obtained based on permutations of $\hat{\bar{u}}$ as described in Section 2.2. The formal properties of this test follow from the results in Section 3.

A.3 Placebo Tests

Here we propose easy-to-implement placebo tests for assessing the credibility of inferences based on our method. We recommend applying these placebo tests when using our inference procedures.

Following Abadie et al. (2015), the idea is to consider a placebo intervention before the actual intervention took place. For a given $\tau \geq 1$, we use our method to test the null hypothesis

$$H_0 : \theta_{T_0-\tau+1} = \cdots = \theta_{T_0} = 0 \tag{A.3}$$

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

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

43

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. The theoretical properties of such placebo tests follow directly from the results in Section 3. As illustrated in our application, it may be useful to complement the formal testing results with plots of the pre-treatment residuals $(\hat{u}_1, \ldots, \hat{u}_{T_0})$.

A rejection of the null hypothesis (A.3) undermines the credibility of the assumptions underlying our procedure and the inferences on the policy effects in the post-treatment period. While non-rejections provide evidence in favor of our method, it is important to emphasize that such non-rejections do not "prove" that our method is valid. Moreover, by construction, the placebo tests cannot be used to assess some of the key assumptions underlying our approach, such as the invariance of the distribution of $\{u_t\}$ under the intervention.

Given our inference method's genericness, the placebo tests discussed here can be used to assess and compare the credibility of different CSC methods. For example, in our empirical application, the placebo tests provide evidence in favor of SC and constrained Lasso, but suggest that the difference-in-differences results need to be interpreted with caution.

B   Interpretation as a Structural Breaks Test

A key assumption underlying our method is the invariance of the distribution of $\{u_t\}$ under the intervention (Assumption 1). In settings where this assumption fails, following the literature on end-of-sample structural breaks tests (e.g., Andrews, 2003), our procedure can be used as a test of the null hypothesis that the policy has no impact whatsoever against the alternative hypothesis that $\theta \neq 0$ and or the policy affects the distribution of $\{u_t\}$.

Consider the following testing problem:

$$
H_0 : \begin{cases} Y_{1t}^N = P_t^N + u_t, \quad E(u_t) = 0, \quad t = 1, \ldots, T, \quad \theta = 0, \quad \text{and} \\ \{u_t\}_{t=0}^T \text{ is stationary and weakly dependent.} \end{cases} \tag{B.1}
$$

against

$$
H_1 : \begin{cases} Y_{1t}^N = P_t^N + u_t, \quad t = 1, \ldots, T, \quad \theta \neq 0 \text{ and or the distribution of} \\ \{u_t\}_{t=T_0+1}^T \text{ differs from that of } \{u_t\}_{t=s}^{s+T_*-1} \text{ for } s = 1, \ldots, T_0 - T_* + 1. \end{cases} \tag{B.2}
$$

Hypothesis (B.1) can be tested by applying our procedure to the data under the null where $Y_{1t}^N = Y_{1t}^I = Y_{1t}$ for all $t = 1, \ldots, T$. The theoretical size properties of this test follow directly from the results in Section 3. This test has power against location shifts induced by $\theta \neq 0$

as well as changes in the distribution of $\{u_t\}$ that increase the quantiles of $S(u)$ (e.g., scale shifts); see Andrews (2003, Section 2.5) for a related discussion.

C   Prediction Sets for Random Policy Effects

In the main text, we assume that the policy effect sequence $\{\theta_t\}$ is fixed (Assumption 1). Here we show that our procedure generates valid prediction sets for θ_t when θ_t is assumed to be random, as, for example, in Cattaneo et al. (2021).[24] Our analysis here is in the same spirit as the classical conformal prediction literature, which aims at constructing prediction intervals for future values of a (random) target quantity of interest. Specifically, we will show that when θ_t is random, Algorithm 1 provides unconditionally valid $(1 - \alpha)$-prediction sets with non-asymptotic performance guarantees.

To state the result, define $\mathbf{Z}^* = (Z_1^*, \ldots, Z_T^*)'$ with $Z_t^* = \left(Y_{1t}^N, Y_{2t}^N, \ldots, Y_{J+1t}^N, X_{1t}', \ldots, X_{J+1t}'\right)'$ for $1 \leq t \leq T_0 + T_*$. Notice that $\mathbf{Z}^*$ contains the true counterfactuals.

Theorem C.1 (Prediction Sets). *Assume that T_* is fixed. Suppose that $Y_{1t}^N = P_t^N + u_t$, $1 \leq t \leq T$, where $\{u_t\}$ is a centered and stationary stochastic process, and that the policy effects $\theta_t := Y_{1t}^I - Y_{1t}^N$ are random. Suppose that Assumption 3 holds for $\hat{P}_N$ computed using $\mathbf{Z}^*$. Impose Assumption 2.1 if $\Pi = \Pi_{all}$. Impose Assumption 2.2, if $\Pi = \Pi_{\rightarrow}$. Assume the statistic $S(u)$ has a density function bounded by D. Then*

$$|P\left(\theta_t \in \mathcal{C}_{1-\alpha}(t)\right) - (1 - \alpha)| \leq C(\tilde{\delta}_T + \delta_T + \sqrt{\delta_T} + \gamma_T),$$

where $\mathcal{C}_{1-\alpha}(t)$ is defined in Algorithm 1 and $\tilde{\delta}_T = (T_*/T_0)^{1/4}(\log T)$. The constant C depends on T_*, M and D, but not on T.

Let us briefly discuss the interpretation of our model when θ_t is regarded as random. Suppose that the potential outcomes are $Y_{1t}^I = P_t^I + u_t^I$ and $Y_{1t}^N = P_t^N + u_t^N$. The counterfactual mean proxies and the prediction errors under the policy, $\{P_t^I\}$ and $\{u_t^I\}$, may differ from the counterfactual mean proxies and the errors in the absence of the policy, $\{P_t^N\}$ and $\{u_t^N\}$. Algorithm 1 provides prediction intervals for

$$\theta_t := Y_{1t}^I - Y_{1t}^N = \Delta_t^P + \Delta_t^u,$$

---

[24]Following the literature on conformal prediction (e.g., Lei and Wasserman, 2014; Lei et al., 2018), we use the terminology "prediction set" instead of "confidence set" because the resulting sets are not conventional confidence sets; see Cattaneo et al. (2021) for a discussion of the differences between confidence intervals and prediction intervals for SC.

where $\Delta_t^P := P_t^I - P_t^N$ and $\Delta_t^u := u_t^I - u_t^N$. Here θ_t captures both the effects of the policy on the mean proxy, Δ_t^P, and the effect on distribution of the error, Δ_t^u (e.g., a scale shift). Therefore, we are only assuming that in the absence of the policy, $\{u_t^N\}$ is a stationary process. All the changes to the mean, the variance, or other features of the distribution due to the policy are captured by the policy effects $\{\theta_t\}$.

In sum, with random policy effects, our procedure simply provides a prediction set for the policy effects defined as the difference between the two potential outcomes.

D   Model-free Exact Validity under Exchangeability

Every permutation procedure that is approximately valid in time series settings should have good properties in "ideal" settings where the data are iid or exchangeable. The following theorem, which is based on standard arguments, shows that under exchangeability of the data, our conformal inference approach achieves exact finite sample size control. This result is model-free in the sense that we do not need to use a correct or consistent estimator for the counterfactual mean proxy. As a result, our procedure controls size under arbitrary forms of misspecification and is fully robust against overfitting.

Theorem D.1 (Exact Validity). *Let Π be $\Pi_\rightarrow$ or Π_{all}. Suppose that the data $\{Z_t\}_{t=1}^T$ are iid or exchangeable with respect to Π under the null hypothesis and 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, under the null hypothesis, $\{\hat{u}_t\}_{t=1}^T$ is an exchangeable sequence and 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 distribution,

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

Theorem D.1 requires that the estimators are invariant under permutations of the data under the null hypothesis. Invariance holds for regression-based estimators such as SC, constrained Lasso, or penalized regression, provided that the null hypothesis is imposed for estimation but may fail for dynamic models such as linear and non-linear autoregressive models.

An inspection of the proof of Theorem D.1 shows that our procedure achieves finite sample size control whenever the residuals $\{\hat{u}_t\}_{t=1}^T$ are exchangeable. We demonstrate that exchangeability of $\{\hat{u}_t\}_{t=1}^T$ is implied if the data $\{Z_t\}_{t=1}^T$ are iid or exchangeable. Exchange-

ability $\{\hat{u}_t\}_{t=1}^T$ may hold even if the data are not exchangeable. For example, in the difference-in-difference model, the outcome data can have an arbitrary common trend eliminated by differencing, making it possible for the residuals to be iid or exchangeable with non-iid data.

The finite sample validity under exchangeability is crucial for the robustness of our proposal. While high-dimensional CSC approaches may overfit in small samples, Theorem D.1 shows that the proposed method does not suffer from overfitting or misspecification. The fundamental reason is that we exploit symmetry rather than to completely rely on the consistency of the estimator. Imposing the null hypothesis when estimating the model leads to invariance of the estimator under permutations, exchangeable residuals, and finite sample validity. As a result, the proposed procedure is more robust than alternative approaches that estimate the proxies based on the pre-treatment data without imposing the null hypothesis.

Even in time series settings where exchangeability fails, imposing the null hypothesis for estimation is crucial for achieving a good performance in typical CSC applications where T_0 is rather small (e.g., Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015; Doudchenko and Imbens, 2016; Cunningham and Shah, 2018). Figure I.1 augments Figure 1 with results for $T = 100$ ($T_0 = 99, T_* = 1$). In the empirically relevant case where $T_0 = 19$, estimating P_t^N under the null yields an excellent performance. Irrespective of the degree of persistence, size accuracy is substantially better than when P_t^N is estimated based on pre-treatment data only. Even when $T_0 = 99$, which is much larger than the T_0 in many CSC applications, imposing the null yields notable performance improvements. In fact, the size accuracy is better for $T_0 = 19$ when P_t^N is estimated under the null than for $T_0 = 99$ when P_t^N is estimated based on the pre-treatment data only.

E   Sufficient Conditions for Estimator Stability

In this section, we provide sufficient conditions for the estimator stability Assumption 4. We first present generic sufficient conditions for low-dimensional models. For high-dimensional models, the theoretical analysis is 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 verify the stability condition for constrained Lasso; stability of Ridge regression is verified in Appendix F.

E.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 E.1. *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 E.1 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. The conclusion of Lemma E.1 translates to Assumption 4 once we derive a bound on $\max_{\pi \in \Pi} \sup_{\beta_1 \neq \beta_2} |S(\mathbf{Z}^\pi; \beta_1) - S(\mathbf{Z}^\pi; \beta_2)| / \|\beta_1 - \beta_2\|_2$; this requires knowledge of the model structure.

For example, suppose that $\max_{\pi \in \Pi} \sup_{\beta_1 \neq \beta_2} |S(\mathbf{Z}^\pi; \beta_1) - S(\mathbf{Z}^\pi; \beta_2)| / \|\beta_1 - \beta_2\|_2 = O_P(1)$.[25] Then $\max_{\pi \in \Pi} |S(\mathbf{Z}^\pi; \hat{\beta}(\mathbf{Z})) - S(\mathbf{Z}^\pi; \hat{\beta}(\mathbf{Z}_H))| = o_P(1)$. Here is how we apply Theorem 2 to obtain the asymptotic size control. Fix an arbitrary $\delta > 0$. We can simply choose the constant function $\varrho_T(x) = \delta$ for Assumption 4. Since $\max_{\pi \in \Pi} |S(\mathbf{Z}^\pi; \hat{\beta}(\mathbf{Z})) - S(\mathbf{Z}^\pi; \hat{\beta}(\mathbf{Z}_H))| = o_P(1)$, Assumption 4 holds for some $\gamma_{1,T} = o(1)$ (due to the definition of convergence in probability). Assume that $\Psi(x, \beta)$ has bounded derivative with respect to x. Then we can choose $\xi_T = K_1$ for a large constant K_1 and $\gamma_{2,T} = 0$. As a result, Theorem 2 has

$$
\begin{aligned}
&|P(\hat{p} \leq \alpha) - \alpha| \\
&\leq C_1 \sqrt{\xi_T \varrho_T (T_0/R + 2k)} + C_1 \left(T_0^{-1} R [\log(T_0/R)]^{1/D_3}\right)^{1/4} + C_1 \exp\left(-(k - T_* + 1)^{1/D_3}\right) \\
&\quad + C_1 \sqrt{\gamma_{1,T}} + C_1 \sqrt{\gamma_{2,T}} \\
&= C_1 \sqrt{K_1 \delta} + C_1 \left(T_0^{-1} R [\log(T_0/R)]^{1/D_3}\right)^{1/4} + C_1 \exp\left(-(k - T_* + 1)^{1/D_3}\right) + C_1 \sqrt{o(1)}.
\end{aligned}
$$

Since $R \asymp T_0 / \log(T_0)$, we have $T_0^{-1} R [\log(T_0/R)]^{1/D_3} = o(1)$. Choosing $k \asymp \log(T_0)$ and assuming that T_* is fixed, we obtain $\exp\left(-(k - T_* + 1)^{1/D_3}\right) = o(1)$. Therefore, the above

---

[25] We only need $S(\mathbf{Z}; \beta)$ to be Lipschitz with respect to β. In the simple example of $S(\mathbf{Z}; \beta) = |Y_{T_0+1} - X'_{T_0+1}\beta|$, this only requires $\|X_{T_0+1}\|_2$ to be bounded.

display implies

$$|P\left(\hat{p} \leq \alpha\right) - \alpha| = C_1\sqrt{K_1\delta} + o(1).$$

Since $\delta > 0$ is arbitrary, we have $|P\left(\hat{p} \leq \alpha\right) - \alpha| = o_P(1)$.

E.2 Constrained Lasso

Here we propose sufficient conditions for estimator stability for constrained Lasso. In contrast to Sections 2.3.2 and 4.2, 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 non-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 iid 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 $H \subset \{1, \ldots, T\}$, let $\hat{\Sigma}_H = T^{-1}\sum_{t=1}^T X_{t,H}X_{t,H}'$ and $\hat{\mu}_H = T^{-1}\sum_{t=1}^T X_{t,H}Y_{t,H}$. Finally, let $\mathcal{H}$ be a set of subsets of $\{1, \ldots, T\}$.

Lemma E.2. *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 E.2 provides sufficient conditions for perturbation stability. Inspecting the proof, we notice that the argument does not require the estimator to converge to anything. To 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 E.2 allows the model to be misspecified and the pseudo-true value may or may not be consistently estimated by constrained Lasso.

Lemma E.2 says that when the solution of constrained Lasso is sparse, the stability of $\hat{\Sigma}$ and $\hat{\mu}$ guarantees 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 E.2 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 $J \ll T_0/\log(T_0)$.

We now give an explicit formula for $\varrho_T(\cdot)$ in Assumption 4. Suppose that $S(\mathbf{Z}, \beta) = |T_*^{-1}\sum_{t=T_0+1}^{T_0+T_*}(Y_t - X'_t\beta)|$. Then the above display implies that $\max_H |S(\mathbf{Z}; \hat{\beta}(\mathbf{Z})) - S(\mathbf{Z}; \hat{\beta}(\mathbf{Z}_H))| = O_P(\sqrt{sT_0^{-1}\log T_0})$. As in Appendix E.1, we can use Theorem 2 to obtain the asymptotic size control. Assume $s = o(T_0/\log(T_0))$. Let $q_T = (sT_0^{-1}\log(T_0))^{-1/4}$. Notice that $q_T \to \infty$. Since $q_T \to \infty$ and $|H| \lesssim \log(T_0)$, the above result of $\max_H |S(\mathbf{Z}; \hat{\beta}(\mathbf{Z})) - S(\mathbf{Z}; \hat{\beta}(\mathbf{Z}_H))| = O_P(\sqrt{sT_0^{-1}\log T_0})$ implies that there exists $\gamma_{1,T} = o(1)$ such that Assumption 4 holds with

$\varrho_T(x) = q_T\sqrt{sT_0^{-1}x}$. Then by the same argument as in Appendix E.1, Theorem 2 implies

$$|P(\hat{p} \leq \alpha) - \alpha| \leq C_1\sqrt{K_1\varrho_T(T_0/R + 2k)} + o(1),$$

where K_1 is the same constant as in Appendix E.1. Since $k \leq T_0/R$, $R \asymp T_0/\log(T_0)$ and $q_T \to \infty$, we have

$$\varrho_T(T_0/R + 2k) \leq \varrho_T(3T_0/R) = \sqrt{3}q_T\sqrt{sT_0^{-1}(T_0/R)} \lesssim q_T\sqrt{sT_0^{-1}\log(T_0)} = 1/q_T \to 0.$$

Remark 3. Stability does not imply that the constrained Lasso residuals, $\hat{\varepsilon}_t$, are close to $\varepsilon_{*,t} = Y_t - X_t'\beta_*$, where $\beta_* = \arg\min_{\beta\in\mathcal{W}} E(Y_t - X_t'\beta)^2$ is a pseudo-true value. In Chernozhukov et al. (2019), we show that $\|\hat{\beta} - \beta_*\|_2 = O_P((T_0^{-1}\log J)^{1/4})$. However, this is far from enough to conclude that $|X_t'(\hat{\beta} - \beta_*)| = o_P(1)$ due to the high-dimensionality of X_t. The usual Cauchy-Schwarz bound $\|X_t\|_2\|\hat{\beta} - \beta_*\|_2$ would not converge to zero; the Hölder bound $\|X_t\|_\infty\|\hat{\beta} - \beta_*\|_1$ does not suffices either since $\|\hat{\beta} - \beta_*\|_1$ does not converge to zero. [When $\mathcal{W}$ is a bounded ℓ_1-ball, it is in fact impossible to achieve $\|\hat{\beta} - \beta_*\|_1 = o_P(1)$ (Ye and Zhang, 2010).] Even if β_* is assumed to be sparse, one would still require $\|\beta_*\|_0 = o(\sqrt{T_0}/\log J)$. In contrast, our stability condition discussed above only requires the much weaker condition of $\|\beta_*\|_0 = o(T_0/\log T_0)$. Therefore, stability condition could be satisfied even if the estimated residual does not converge to a pseudo-true target. $\square$

F   Consistency and Estimator Stability

In this section, we illustrate the difference between stability and consistency using a simple and analytically tractable example: Ridge regression. We shall show that under correct specification, Ridge may not be consistent, while still satisfying estimator stability.

To keep theoretical analysis tractable, we work under stylized conditions. Given data $Z = (Y, X)$, we define the ridge estimator

$$\hat{\beta}_\lambda(Z) = (X'X + \lambda I_T)^{-1}X'Y,$$

where λ is a tuning parameter. Then we can compute the residuals $\hat{u}(Z; \lambda) = Y - X\hat{\beta}_\lambda(Z)$.

Lemma F.1. *Suppose that $Y = X\beta + u$, where $X = (X_1, \ldots, X_T)' \in \mathbb{R}^{T\times J}$ and $E(u \mid X) = 0$. Assume that $\|\beta\|_2$ is bounded away from zero and infinity, $E(uu' \mid X) = \sigma^2 I_T$ for a constant $\sigma > 0$, $J \lesssim T^{\kappa_0}$ with $\kappa_0 \in (0, 2/3)$ and for some constants $\kappa_1, \kappa_2 > 0$,*

$$P(\kappa_1 T \leq \lambda_{\min}(X'X) \leq \lambda_{\max}(X'X) \leq \kappa_2 T) \geq 1 - o(1)$$

and

$$P\left(\max_{1 \leq t \leq T} \|X_t\| \leq \kappa_3 \sqrt{J} \quad \text{and} \quad \|X'u\|_2 \leq \kappa_4 \sqrt{JT}\right) \geq 1 - o(1).$$

Let $Z_H = (\tilde{Y}, \tilde{X})$ be the perturbed data with $|H| \asymp \log T$. For $\tilde{u} = \tilde{Y} - \tilde{X}\beta$, assume that

$$P\left(\|\tilde{X}'\tilde{u} - X'u\|_2 \leq \kappa_5 \sqrt{J|H|}\right) \geq 1 - o(1)$$

and

$$P\left(\|\tilde{X}'\tilde{X} - X'X\| \leq \kappa_6(|H| + J)\right) \geq 1 - o(1),$$

where $\kappa_5, \kappa_6 > 0$ are constants. If $\lambda \asymp T$, then $E\|X(\hat{\beta}_\lambda(Z) - \beta)\|_2^2 \gtrsim T$ and $\|\hat{u}(Z; \lambda) - \hat{u}(Z_H; \lambda)\|_\infty = o_P(1)$[26].

Lemma F.1 provides a robustness guarantee for the validity of the procedure. Under the ideal choice of the tuning parameter λ, we would expect consistency of $\hat{\beta}_\lambda(Z)$ and hence validity of the procedure. However, Lemma F.1 states that even when the tuning parameter is badly chosen such that consistency fails, one might still expect the estimator to be stable under perturbations, which is sufficient for the validity of our inference procedure. This is important in practice since computing optimal tuning parameters is often difficult.

Now we give very simple sufficient (but probably far from necessary) conditions for the assumptions of Lemma F.1. Suppose that the data is independent across t and rows of X are sub-Gaussian. Then standard results in random matrix theory can be used to verify these assumptions. For example, $X'X/T - E(X'X/T)$ has eigenvalues tending to zero (e.g., Theorem 5.39 and Remark 5.40 in Vershynin (2010)). Hence, as long as $E(X'X/T)$ has bounded eigenvalues, there exist constants κ_1, κ_2 that satisfy the condition in Lemma F.1. Let us further assume that entries of X are also bounded and entries of u are sub-Gaussian and independent of X. Then κ_3 can be chosen to be a large enough constant.

The existence of κ_4, κ_5 is a consequence of the Hanson-Wright inequality. We consider $\|X'u\|_2$ conditional on X. By Theorem 6.3.2 in Vershynin (2018) (proved using the Hanson-Wright inequality) applied to the conditional probability given X and by the definition of sub-Gaussian variables, we have that with probability one, for any $z > 0$,

$$P\left(\|X'u\|_2 - \sigma\|X\|_F| > z\|X\| \mid X\right) \leq C_1 \exp\left(-C_2 z^2\right),$$

where $C_1, C_2 > 0$ are constants depending only on the sub-Gaussian norm of u. Choosing

---

[26]Notice that consistency is enough to derive asymptotic size control from Theorem 2; see the discussion at the end of Appendix E.1 for an explicit formula for quantities in Assumption 4.

$z = \sqrt{\log(T)}$, we obtain

$$P\left(\|X'u\|_2 \leq \sigma\|X\|_F + \sqrt{\log(T)}\|X\|\right) \leq o(1).$$

Notice that $E\|X\|_F^2 = \sum_{j,t} EX_{j,t}^2 = O(JT)$ and $\|X\|_F^2 - E\|X\|_F^2 = \sum_t(\sum_j(X_{j,t}^2 - EX_{j,t}^2))$. By the boundedness of $X_{j,t}$, $\sum_j(X_{j,t}^2 - EX_{j,t}^2) = O(J)$ has mean zero and variance bounded by $O(J^2)$. It follows by the central limit theorem that $\|X\|_F^2 - E\|X\|_F^2 = \sum_t(\sum_j(X_{j,t}^2 - EX_{j,t}^2)) = \sqrt{T}J$. Since $E\|X\|_F^2 = O(JT)$, there exists a constant $C_3 > 0$ such that $P(\|X\|_F^2 < C_3JT) \geq 1 - o(1)$. Again by the random matrix theory (e.g., Theorem 5.39 and Remark 5.40 in Vershynin (2010)), $\|X\| = O_P(\sqrt{J+T})$. We have that with probability approaching one, $\|X'u\|_2 \leq \sigma\|X\|_F + \sqrt{\log(T)}\|X\| \leq \sqrt{C_3JT}\sigma + O(\sqrt{(J+T)\log(T)})$. Assume $J \gg \log(T)$. Since $(J+T)\log(T) \ll JT$, there exists κ_4 satisfying $P(\|X'u\|_2 \leq \kappa_4\sqrt{JT}) \geq 1 - o(1)$. The same argument holds for κ_5; we just repeat the same argument with T replaced by $|H|$ once we realize $\tilde{X}'\tilde{u} - X'u = \sum_{t \in H} \tilde{X}_t\tilde{u}_t - \sum_{t \in H} X_t u_t$, where X_t denotes the t-th row of X. (Analogous notations apply to $\tilde{X}_t, u_t$, etc.)

To find κ_6, we observe that $\tilde{X}'\tilde{X} - X'X = \sum_{t \in H} \tilde{X}_t\tilde{X}_t' - \sum_{t \in H} X_t X_t'$. By the random matrix theory (e.g., Theorem 5.39 and Remark 5.40 in Vershynin (2010)), $\|\sum_{t \in H}(X_t X_t' - EX_t X_t')\| \leq C_4\sqrt{J|H|}$ with probability approaching one, where $C_4 > 0$ is a constant. Again assuming that $EX_t X_t'$ has eigenvalues bounded by a constant $C_5 > 0$, we have that $\|\sum_{t \in H} X_t X_t'\| \leq C_5|H| + C_4\sqrt{J|H|}$ with probability approaching one. Recall the elementary inequality $\sqrt{J|H|} \leq (J + |H|)/2$. Therefore, we can simply set $\kappa_6 = 4(C_4 + C_5)$ and obtain that with probability approaching one, $\|\sum_{t \in H} X_t X_t'\| \leq (|H| + J)\kappa_6/2$. Similarly, we can get a bound for $\|\sum_{t \in H} \tilde{X}_t\tilde{X}_t'\|$.

The above analysis is based on simple sub-Gaussian or boundedness assumptions on (X, u). However, the purpose is to show that even in the simple case, Ridge regression can exhibit stability without being consistent (due to Lemma F.1). We expect similar results to hold in more general or complicated data-generating processes. We leave this extension for future research.

G   Simulation Study

This section presents simulation evidence on the finite sample properties of our inference procedures. We consider the three CSC methods used in the empirical application in Section 5: difference-in-differences, canonical SC, and constrained Lasso with $K = 1$.

We consider different data generating processes (DGPs) for the treated unit all of which

specify the treated outcome as a weighted combination of the control outcomes:

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

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 (2017), the control outcomes are generated using a factor model:

$$Y_{jt}^N = \lambda_{1j} + F_{1t} + \lambda_{2j} F_{2t} + \epsilon_{jt},$$

where $\lambda_{1j} = (j-1)/J$, $\lambda_{2j} = (j-1)/J$, $F_{1t} \overset{iid}{\sim} N(0,1)$, 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 vary ρ_u, ρ_ϵ, T_0, J, and F_{2t}. The DGPs differ with respect to the specification of the weights w.

	Weight Specification	Correctly Specified Model(s)
DGP1	$w = \left(\frac{1}{J}, \ldots, \frac{1}{J}\right)'$	Difference-in-differences, SC, constrained Lasso
DGP2	$w = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, \ldots, 0\right)'$	SC, constrained Lasso
DGP3	$w = -1 \cdot \left(\frac{1}{J}, \ldots, \frac{1}{J}\right)'$	constrained Lasso
DGP4	$w = (1, -1, 0, \ldots, 0)'$	–

We set $T_* = 1$ and consider the problem of testing the null hypothesis of a zero effect:

$$H_0 : \theta_T = 0.$$

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

We first analyze the performance of our procedure with stationary data, letting $F_{2t} \overset{iid}{\sim} N(0,1)$. Table I.1 presents simulation evidence on the size properties of our method when the data are iid ($\rho_u = \rho_\epsilon = 0$), which implies exchangeability of the residuals (cf. Theorem D.1). As expected, our procedure achieves exact size control, irrespective of whether or not the model for P_t^N is correctly specified. To study the finite sample performance with dependent data, we set $\rho_u = \rho_\epsilon = 0.6$. Table I.2 shows that our method exhibits close-to-correct size under correct specification and under misspecification, confirming the theoretical results on the robustness of our procedure under estimator stability and stationarity.

To investigate the performance with non-stationary data, we use trending factors: $F_{2t} \sim N(t, 1)$. Tables I.3 and I.4 show that under correct specification, our method exhibits excellent size properties. However, unlike in stationary settings, misspecification can cause size distortions. This finding is expected given the theoretical results and discussions in Section

Figure I.2 displays power curves for a setting where $T_0 = 19$ and $J = 50$ as in our empirical application, $\rho_u = \rho_\epsilon = 0.6$, and $F_{2t} \overset{iid}{\sim} N(0,1)$. Under correct specification, our method exhibits excellent small sample power properties and comes close to achieving the oracle power bound based on the true marginal distribution of u_t. Moreover, we find that imposing additional constraints when estimating P_t^N (e.g., using SC instead of the more general constrained Lasso) does not improve power when these additional restrictions are correct but can cause power losses when they are not.

H   Proofs

Additional Notation

We introduce some additional notations that will be used in the proofs. For $a, b \in \mathbb{R}$, $a \vee b = \max\{a, b\}$. For two positive sequences a_n, b_n (indexed by 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.

H.1   Proof of Theorem 1

The proof proceeds by verifying the high-level conditions in the following lemma. Let $n = |\Pi|$ so that $n = T!$ if $\Pi = \Pi_{\mathrm{all}}$ and $n = T$ if $\Pi = \Pi_{\rightarrow}$.

Lemma H.1 (Approximate Validity under High-Level Conditions[27]). *Let $\{\delta_{1n}, \delta_{2n}, \gamma_{1n}, \gamma_{2n}\}$ be sequences of numbers converging to zero. 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 $F(x) = P(S(u) < x)$, namely

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

---

[27]In Chernozhukov et al. (2018), we use a version of Lemma H.1, 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. The proof of Lemma H.1 (presented in Section H.1.1 to make the exposition self-contained) follows from the same arguments as the proof of Theorem 2 in Chernozhukov et al. (2018) modified to the problem of permuting residuals.

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

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

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

      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 p-value obeys for any $\alpha \in (0, 1)$

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

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 H.1. 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 H.2 (Approximate Ergodicity under Moving Block Permutations). *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) $T_0 \ge T_* + 2$, and (3) $S(u)$ has bounded pdf. Then there exists a constant $M' > 0$ depending only on M such that for any $\delta_{1n} > 0$,*

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

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 iid permutations.

Lemma H.3 (Approximate Ergodicity under iid Permutations). *Let Π be the set of all permutations. Suppose that $\{u_t\}_{t=1}^T$ is iid. Assume that $S(u)$ only depends on the last T_* entries of u. If $T_0 \ge T_* + 2$, then*

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

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 H.4 (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 $1 - \gamma_T$,*

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

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

Lemma H.5 (Bounds on Estimation Errors under iid 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 $1 - \gamma_T$,*

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

Now we conclude the proof of Theorem 1.

For the moving block permutations, let $\delta_{1n} = (T_*/T_0)^{1/4}$. Then we apply Lemma H.1 together with Lemmas H.2 and H.4, obtaining

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

$$\leq 4\delta_{1n} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}}) + \left(M'\sqrt{\frac{T_*}{T_0}}\log T_0 + \frac{T_* + 1}{T_0 + T_*}\right)/\delta_{1n} + \gamma_{2n}$$

$$\leq 4(T_*/T_0)^{1/4} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}})$$

$$+ \left(M'\sqrt{\frac{T_*}{T_0}}\log T_0 + \frac{T_* + 1}{T_0 + T_*}\right)(T_*/T_0)^{-1/4} + \gamma_{2n}.$$

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 iid permutations, we also use $\delta_{1n} = (T_*/T_0)^{1/4}$. Then we apply Lemma H.1 together with Lemmas H.3 and H.5, obtaining

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

$$\leq 4\delta_{1n} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}}) + \sqrt{2\pi/\lfloor T/T_* \rfloor}/\delta_{1n} + \gamma_{2n}$$
$$\leq 4(T_*/T_0)^{1/4} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}}) + \sqrt{2\pi/\lfloor T/T_* \rfloor}(T_*/T_0)^{-1/4} + \gamma_{2n}$$
$$\lesssim (T_*/T_0)^{1/4} + \delta_{2n} + \sqrt{\delta_{2n}} + \gamma_{2n}.$$

This completes the proof for iid permutations.

H.1.1 Proof of Lemma H.1

The proof proceeds in two steps.[28]

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,

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

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

$$\left| \hat{F}(S(\hat{u})) - F(S(u)) \right| \leq \left| \hat{F}(S(\hat{u})) - F(S(\hat{u})) \right| + |F(S(\hat{u})) - F(S(u))|$$
$$\overset{(i)}{\leq} \sup_{x \in \mathbb{R}} \left| \hat{F}(x) - F(x) \right| + D |S(\hat{u}) - S(u)|$$
$$\leq \sup_{x \in \mathbb{R}} \left| \hat{F}(x) - \tilde{F}(x) \right| + \sup_{x \in \mathbb{R}} \left| \tilde{F}(x) - F(x) \right| + D |S(\hat{u}) - S(u)|$$
$$\leq \sup_{x \in \mathbb{R}} \left| \hat{F}(x) - \tilde{F}(x) \right| + \delta_{1n} + D |S(\hat{u}) - S(u)|$$
$$\leq \sup_{x \in \mathbb{R}} \left| \hat{F}(x) - \tilde{F}(x) \right| + \delta_{1n} + D\delta_{2n}, \tag{H.2}$$

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 $x \in \mathbb{R}$,

$$\left| \hat{F}(x) - \tilde{F}(x) \right|$$
$$\leq \frac{1}{n} \sum_{\pi \in A} |\mathbf{1}\{S(\hat{u}_\pi) < x\} - \mathbf{1}\{S(u_\pi) < x\}| + \frac{1}{n} \sum_{\pi \in (\Pi \setminus A)} |\mathbf{1}\{S(\hat{u}_\pi) < x\} - \mathbf{1}\{S(u_\pi) < x\}|$$

---

[28]The proof follows from the same arguments as the proof of Theorem 2 in Chernozhukov et al. (2018) modified to the problem of permuting residuals and is presented for completeness.

58

$$\overset{(i)}{\leq} \frac{|A|}{n} + \frac{1}{n} \sum_{\pi \in (\Pi \backslash A)} \mathbf{1} \left\{ |S(u_\pi) - x| \leq \sqrt{\delta_{2n}} \right\} \leq \frac{|A|}{n} + \frac{1}{n} \sum_{\pi \in \Pi} \mathbf{1} \left\{ |S(u_\pi) - x| \leq \sqrt{\delta_{2n}} \right\}$$

$$\leq \frac{|A|}{n} + P\left(|S(u) - x| \leq \sqrt{\delta_{2n}} \right)$$

$$+ \sup_{z \in \mathbb{R}} \left| \frac{1}{n} \sum_{\pi \in \Pi} \mathbf{1} \left\{ |S(u_\pi) - z| \leq \sqrt{\delta_{2n}} \right\} - P\left(|S(u) - z| \leq \sqrt{\delta_{2n}} \right) \right|$$

$$= \frac{|A|}{n} + P\left(|S(u) - x| \leq \sqrt{\delta_{2n}} \right)$$

$$+ \sup_{x \in \mathbb{R}} \left| \left[\tilde{F}\left(z + \sqrt{\delta_{2n}} \right) - \tilde{F}\left(z - \sqrt{\delta_{2n}} \right) \right] - \left[F\left(z + \sqrt{\delta_{2n}} \right) - F\left(z - \sqrt{\delta_{2n}} \right) \right] \right|$$

$$\leq \frac{|A|}{n} + P\left(|S(u) - x| \leq \sqrt{\delta_{2n}} \right) + 2 \sup_{z \in \mathbb{R}} \left| \tilde{F}(z) - F(z) \right|$$

$$\overset{(ii)}{\leq} \frac{|A|}{n} + 2D\sqrt{\delta_{2n}} + 2\delta_{1n} \overset{(iii)}{\leq} \delta_{1n} + 2\delta_{2n} + 2D\sqrt{\delta_{2n}}, \tag{H.3}$$

where (i) follows by 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 $\mathcal{M}$,

$$\sup_{x \in \mathbb{R}} \left| \hat{F}(x) - \tilde{F}(x) \right| \leq \delta_{1n} + 2\delta_{2n} + 2D\sqrt{\delta_{2n}}. \tag{H.4}$$

We combine (H.2) and (H.4) and obtain that on the event $\mathcal{M}$,

$$\left| \hat{F}(S(\hat{u})) - F(S(u)) \right| \leq 2\delta_{1n} + 2\delta_{2n} + D(\delta_{2n} + 2\sqrt{\delta_{2n}}). \tag{H.5}$$

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

$$\left| P\left(1 - \hat{F}(S(\hat{u})) \leq \alpha \right) - \alpha \right|$$

$$= \left| E\left(\mathbf{1}\left\{ 1 - \hat{F}(S(\hat{u})) \leq \alpha \right\} - \mathbf{1}\left\{ 1 - F(S(u)) \leq \alpha \right\} \right) \right|$$

$$\leq E\left| \mathbf{1}\left\{ 1 - \hat{F}(S(\hat{u})) \leq \alpha \right\} - \mathbf{1}\left\{ 1 - F(S(u)) \leq \alpha \right\} \right|$$

$$\overset{(i)}{\leq} P\left(|F(S(u)) - 1 + \alpha| \leq \left| \hat{F}(S(\hat{u})) - F(S(u)) \right| \right)$$

$$\leq P\left(|F(S(u)) - 1 + \alpha| \leq \left| \hat{F}(S(\hat{u})) - F(S(u)) \right| \text{ and } \mathcal{M} \right) + P(\mathcal{M}^c)$$

$$\overset{(ii)}{\leq} P\left(|F(S(u)) - 1 + \alpha| \leq 2\delta_{1n} + 2\delta_{2n} + D(\delta_{2n} + 2\sqrt{\delta_{2n}}) \right) + P(\mathcal{M}^c)$$

$$\overset{(iii)}{\leq} 4\delta_{1n} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}}) + \gamma_{1n} + \gamma_{2n},$$

where (i) follows by the elementary inequality $|\mathbf{1}\{1-\hat{F}(S(\hat{u}))\le\alpha\}-\mathbf{1}\{1-F(S(u))\le\alpha\}|\le$ $\mathbf{1}\{|F(S(u))-1+\alpha|\le|\hat{F}(S(\hat{u}))-F(S(u))|\}$, (ii) follows by (H.5), (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 (H.1). The proof is complete.

H.1.2   Proof of Lemma H.2

We define

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

It is straight-forward to verify that

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

Let $\tilde{\alpha}_{\mathrm{mixing}}$ 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 $\check{F}(x) = T_0^{-1}\sum_{t=1}^{T_0}\mathbf{1}\{s_t\le x\}$. The bounded pdf of $S(u)$ implies the continuity of $F(\cdot)$. It follows, by Proposition 7.1 of Rio (2017), that

$$
E\left(\sup_{x\in\mathbb{R}}\left|\check{F}(x)-F(x)\right|^2\right) \le \frac{1}{T_0}\left(1+4\sum_{k=0}^{T_0-1}\tilde{\alpha}_{\mathrm{mixing}}(t)\right)\left(3+\frac{\log T_0}{2\log 2}\right)^2. \tag{H.6}
$$

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

$$
\sum_{k=0}^{T_0-1}\tilde{\alpha}_{\mathrm{mixing}}(t) = \sum_{k=0}^{T_*}\tilde{\alpha}_{\mathrm{mixing}}(t) + \sum_{k=T_*+1}^{T_0-1}\tilde{\alpha}_{\mathrm{mixing}}(t) \le 2(T_*+1) + \sum_{k=1}^{T_0-T_*-1}\alpha_{\mathrm{mixing}}(k)
$$

$$
\le 2(T_*+1) + \sum_{k=1}^{\infty}\alpha_{\mathrm{mixing}}(k).
$$

Since $\sum_{k=1}^{\infty}\alpha_{\mathrm{mixing}}(k)$ is bounded by M, it follows by (H.6) 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 $(T_0 + T_*)\tilde{F}(x) - T_0\check{F}(x) = \sum_{t=T_0+1}^{T_0+T_*} \mathbf{1}\{s_t \leq x\}$, it follows that

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

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

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

The desired result follows by Markov's inequality.

H.1.3   Proof of Lemma H.3

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

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

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 $1 \leq i \leq k_T$, 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

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

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

$$E \left(\sup_{x \in \mathbb{R}} \left| \frac{1}{k_T} \sum_{i=1}^{k_T} Q(u_{\pi(b_i)}; x) \right| \right) = \int_0^1 P \left(\sup_{x \in \mathbb{R}} \left| \frac{1}{k_T} \sum_{i=1}^{k_T} Q(u_{\pi(b_i)}; x) \right| > z \right) dz$$

$$\leq \int_0^1 2 \exp \left(-2k_T z^2 \right) dz < \int_0^\infty 2 \exp \left(-2k_T z^2 \right) dz = \sqrt{\pi/(2k_T)},$$

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 iid 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

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

The desired result follows by Markov's inequality.

H.1.4   Proof of Lemma H.4

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

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

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

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.

H.1.5 Proof of Lemma H.5

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 $t \in \{1, ..., T\}$,

$$\sum_{\pi \in \Pi} \left(\hat{u}_{\pi(t)} - u_{\pi(t)}\right)^2 = \sum_{s=1}^{T} \sum_{\pi \in A_{t,s}} \left(\hat{u}_{\pi(t)} - u_{\pi(t)}\right)^2$$

$$= \sum_{s=1}^{T} \sum_{\pi \in A_{t,s}} (\hat{u}_s - u_s)^2 = \sum_{s=1}^{T} |A_{t,s}| (\hat{u}_s - u_s)^2 = (T-1)! \times \|\hat{u} - u\|_2^2.$$

(H.7)

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

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

$$= \frac{Q}{|\Pi|} \sum_{t=T_0+1}^{T_0+T_*} \sum_{\pi \in \Pi} \left(\hat{u}_{\pi(t)} - u_{\pi(t)}\right)^2 = \frac{Q}{|\Pi|} T_*(T-1)! \times \|\hat{u} - u\|_2^2 = QT^{-1}T_*\|\hat{u} - u\|_2^2,$$

where the penultimate equality follows by (H.7) 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.

H.2 Proof of Theorem 2

We first state auxiliary results.

Lemma H.6. *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*

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

The next lemma was derived by Berbee (1987, Lemma 2.1); see Theorem 16.2.1 in Athreya and Lahiri (2006) and Lemma 7.1 of Chen et al. (2016) for popular versions. We state the result to make the exposition more self-contained. The proof is omitted.

Lemma H.7. *Let (W, R) be random vectors defined in the same probability space. Let $P_{W,R}$ denote the probability distribution of (W, R). Let P_R and P_W denote the probability distributions of R and W, respectively. Let $\|\cdot\|_{TV}$ denote the total variation metric. Define the β-mixing coefficient $\beta(W, R) = \|P_{W,R} - P_W \otimes P_R\|_{TV}/2$. Then the probability space can be extended to construct random element $\tilde{W}$ such that (1) $\tilde{W}$ and R are independent, (2) $\tilde{W}$ and W have the same distribution and (3) $P(\tilde{W} \neq W) \leq \beta(W, R)$.*

Now we prove Theorem 2. 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

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

We note that under $\Pi = \Pi_{\rightarrow}$, $\hat{F}(x)$ can be written as

$$\hat{F}(x) = T^{-1}\left(\sum_{t=1}^{T_0}\mathbf{1}\left\{\phi\left(Z_t, ..., Z_{t+T_*-1}; \hat{\beta}(\mathbf{Z})\right) \leq x\right\} + \sum_{t=T_0+1}^{T_0+T_*}\mathbf{1}\left\{\phi\left(Z_{q(t)}, ..., Z_{q(t+T_*-1)}; \hat{\beta}(\mathbf{Z})\right) \leq x\right\}\right),$$
(H.8)

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 $\sup_{x\in\mathbb{R}}|\hat{F}(x) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\tilde{H}_R})\right)|$, 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 $\sup_{x\in\mathbb{R}}\left|\tilde{F}_j(x) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\tilde{H}_j})\right)\right|$.

Let $A_j = \bigcup_{t\in H_j}\{t, ..., t+T_*-1\}$. Since $k > T_*$, we have that $A_j \subset \tilde{H}_j$ and $\min_{t\in A_j, s\in\tilde{H}_j^c}|t-s| \geq k - T_* + 1$. This means that $\{Z_t\}_{t\in\tilde{H}_j^c}$ and $\{Z_t\}_{t\in A_j}$ have a gap of at least $k - T_* + 1$ time

periods. By Lemma H.7 (applied with $W = \{Z_t\}_{t \in \widetilde{H}_j^c}$ and $R = \{Z_t\}_{t \in A_j}$), 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

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

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, $P(\mathcal{M}_j^c) \leq \beta_{\mathrm{mixing}}(k - T_* + 1) + \gamma_{1,T} + \gamma_{2,T}$.

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 $\tilde{\beta}_{\mathrm{mixing}}(i) \leq \beta_{\mathrm{mixing}}(i - T_* + 1)$ for $i \geq T_*$. Moreover, since $(\bar{Z}_t, ..., \bar{Z}_{t+T_*-1})$ is independent of $\mathbf{Z}_{\widetilde{H}_j}$, we have

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

Hence, by Lemma H.6, we have that for any $m_1 \leq m/2$,

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

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

$$
E(a_j) \leq 2m^{1/2} \beta_{\mathrm{mixing}}(m_1 - T_* + 1) + \sqrt{\pi m_1/(2m)} + (m_1 - 1)/m + 2P(\mathcal{M}_j^c), \qquad \text{(H.9)}
$$

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

65

$\mathcal{M}_j$,

$$\left| \tilde{F}_j(x) - \dot{F}_j(x) \right|$$

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

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

$$\overset{(i)}{\leq} m^{-1} \sum_{t \in H_j} \mathbf{1}\left\{ \left| \phi\left(Z_t, ..., Z_{t+T_*-1}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j}) \right) - x \right| \leq \varrho_T(|\widetilde{H}_j|) \right\}$$

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

$$\quad - m^{-1} \sum_{t \in H_j} \mathbf{1}\left\{ \phi\left(Z_t, ..., Z_{t+T_*-1}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j}) \right) < x - \varrho_T(|\widetilde{H}_j|) \right\}$$

$$< \dot{F}_j\left(x + \varrho_T(|\widetilde{H}_j|) \right) - \dot{F}_j\left(x - 2\varrho_T(|\widetilde{H}_j|) \right)$$

$$\leq \Psi\left(x + \varrho_T(|\widetilde{H}_j|); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j}) \right) - \Psi\left(x - 2\varrho_T(|\widetilde{H}_j|); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j}) \right) + 2a_j \overset{(ii)}{\leq} 3\xi_T \varrho_T(|\widetilde{H}_j|) + 2a_j,$$

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 (H.9) and the fact that $\tilde{F}_j(\cdot)$ and $\Psi(\cdot, \cdot)$ take values in $[0, 1]$, we have that for a universal constant $C_1 > 0$,

$$E\left(\sup_{x \in \mathbb{R}} \left| \tilde{F}_j(x) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j}) \right) \right| \right) \tag{H.10}$$

$$\leq 3\xi_T \varrho_T(m + 2k) + 3E(a_j) + 2P(\mathcal{M}_j^c)$$

$$\leq 3\xi_T \varrho_T(m + 2k) + 6m^{1/2}\beta_{\text{mixing}}(m_1 - T_* + 1) + 3\sqrt{\pi m_1/(2m)} + 3(m_1 - 1)/m + 8P(\mathcal{M}_j^c)$$

$$\overset{(i)}{\leq} C_1\left(\xi_T \varrho_T(m + 2k) + m^{1/2}\beta_{\text{mixing}}(m_1 - T_* + 1) + \sqrt{m_1/m} + \beta_{\text{mixing}}(k - T_* + 1) + \gamma_{1,T} + \gamma_{2,T} \right),$$

where (i) follows by $P(\mathcal{M}_j^c) \leq \beta_{\mathrm{mixing}}(k - T_* + 1) + \gamma_{1,T} + \gamma_{2,T}$ 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 $1 \leq j \leq R$,

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

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 $\mathcal{Q}_j$,

$$\left| \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \right|$$
$$= \left| E\left(\mathbf{1}\left\{ \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) \leq x \right\} - \mathbf{1}\left\{ \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \leq x \right\} \mid \mathbf{Z}, \{\tilde{Z}_t\}_{t=1}^{T} \right) \right|$$
$$\leq E\left(\left| \mathbf{1}\left\{ \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) \leq x \right\} - \mathbf{1}\left\{ \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \leq x \right\} \right| \mid \mathbf{Z}, \{\tilde{Z}_t\}_{t=1}^{T} \right)$$
$$\overset{(i)}{\leq} E\left[\mathbf{1}\left\{ \left| \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) - x \right| \right.\right.$$
$$\left.\left. \leq \left| \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) \right| \right\} \mid \mathbf{Z}, \{\tilde{Z}_t\}_{t=1}^{T} \right]$$
$$= P\left[\left| \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) - x \right| \right.$$
$$\left. \leq \left| \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) \right| \mid \mathbf{Z}, \{\tilde{Z}_t\}_{t=1}^{T} \right]$$
$$\leq P\left[\left| \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) - x \right| \leq 2\varrho_T(|\widetilde{H}_j|) \mid \mathbf{Z}, \{\tilde{Z}_t\}_{t=1}^{T} \right]$$
$$+ P\left[\left| \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) \right| > 2\varrho_T(|\widetilde{H}_j|) \mid \mathbf{Z}, \{\tilde{Z}_t\}_{t=1}^{T} \right]$$
$$= \Psi\left(x + 2\varrho_T(|\widetilde{H}_j|); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) - \Psi\left(x - 2\varrho_T(|\widetilde{H}_j|); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)$$

$$+ P\left[\left|\phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\right| > 2\varrho_T(|\widetilde{H}_j|) \mid \mathbf{Z}, \{\tilde{Z}_t\}_{t=1}^T\right]$$

$$\overset{(ii)}{\leq} 4\xi_T \varrho_T(m+2k)$$

$$+ P\left[\left|\phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\right| > 2\varrho_T(|\widetilde{H}_j|) \mid \mathbf{Z}, \{\tilde{Z}_t\}_{t=1}^T\right],$$

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

$$E\left(\sup_{x\in\mathbb{R}} \left|\Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right|\right)$$

$$\leq 4\xi_T \varrho_T(m+2k)$$

$$+ 2P(\mathcal{Q}_j^c) + P\left[\left|\phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\right| > 2\varrho_T(|\widetilde{H}_j|)\right]$$

$$\leq 4\xi_T \varrho_T(m+2k) + 2P(\mathcal{Q}_j^c)$$

$$+ P\left[\left|\phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z})\right)\right| > \varrho_T(|\widetilde{H}_j|)\right]$$

$$+ P\left[\left|\phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) - \phi\left(\dot{Z}_{T_0+1}, ..., \dot{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z})\right)\right| > \varrho_T(|\widetilde{H}_j|)\right]$$

$$\overset{(i)}{\leq} 4\xi_T \varrho_T(m+2k) + 2P(\mathcal{Q}_j^c) + 2\gamma_{1,T} \overset{(ii)}{\leq} 4\xi_T \varrho_T(m+2k) + 2\gamma_{1,T} + 2\gamma_{2,T},$$

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

$$E\left(R^{-1}\sum_{j=1}^{R}\sup_{x\in\mathbb{R}}\left|\Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right|\right) \leq 4\xi_T\varrho_T(m+2k) + 2\gamma_{1,T} + 2\gamma_{2,T}. \quad \text{(H.11)}$$

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 (H.8), we notice that

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

$$\left. + \sum_{t=T_0+1}^{T_0+T_*}\mathbf{1}\left\{\phi\left(Z_{q(t)}, ..., Z_{q(t+T_*-1)}; \hat{\beta}(\mathbf{Z})\right) \leq x\right\}\right| \leq T_* + (T_0 - mR) \leq T_* + R - 1.$$

Moreover, by (H.10) and (H.11), we have that

68

$$E\left(\sup_{x\in\mathbb{R}}\left|R^{-1}\sum_{j=1}^{R}\tilde{F}_j(x) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\tilde{H}_R})\right)\right|\right)$$

$$\leq C_2\left(\xi_T\varrho_T(m+2k) + m^{1/2}\beta_{\text{mixing}}(m_1 - T_* + 1) + \sqrt{m_1/m} + \beta_{\text{mixing}}(k - T_* + 1) + \gamma_{1,T} + \gamma_{2,T}\right)$$

for some universal constant $C_2 > 0$.

The above two displays imply that

$$\sup_{x\in\mathbb{R}}\left|\frac{T}{mR}\hat{F}(x) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\tilde{H}_R})\right)\right| \leq \frac{T_* + R - 1}{mR}$$

$$+C_2\left(\xi_T\varrho_T(m+2k) + m^{1/2}\beta_{\text{mixing}}(m_1 - T_* + 1) + \sqrt{m_1/m} + \beta_{\text{mixing}}(k - T_* + 1) + \gamma_{1,T} + \gamma_{2,T}\right).$$

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 $C_3 > 0$,

$$E\left(\sup_{x\in\mathbb{R}}\left|\hat{F}(x) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\tilde{H}_R})\right)\right|\right)$$

$$\leq C_3\left(\xi_T\varrho_T(m+2k) + m^{1/2}\beta_{\text{mixing}}(m_1 - T_* + 1) + \sqrt{m_1/m} + \beta_{\text{mixing}}(k - T_* + 1) + \gamma_{1,T} + \gamma_{2,T}\right).$$

$$\text{(H.12)}$$

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 $\mathbf{Z}_{\tilde{H}_R}$, (2) $\{\bar{Z}_t\}_{t\in A_R} \stackrel{d}{=} \{Z_t\}_{t\in A_R}$ 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 $\bar{G}(\mathbf{Z}_{\tilde{H}_R}) = \phi\left(\bar{Z}_{T_0+1}, ..., \bar{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\tilde{H}_R})\right)$. 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 $\mathbf{Z}_{\tilde{H}_R}$, which means that

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

Therefore,

conditional on $\hat{\beta}(\mathbf{Z}_{\tilde{H}_R})$, $\Psi\left(\bar{G}(\mathbf{Z}_{\tilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\tilde{H}_R})\right)$ has uniform distribution on $(0, 1)$. (H.13)

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

$$\left| P\left(\hat{F}(G(\mathbf{Z})) < \alpha \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha \right|$$

$$\overset{(i)}{=} \left| E\left(\mathbf{1}\left\{\hat{F}(G(\mathbf{Z})) < \alpha\right\} \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - E\left(\mathbf{1}\left\{\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) < \alpha\right\} \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \right|$$

$$\leq E\left(\left| \mathbf{1}\left\{\hat{F}(G(\mathbf{Z})) < \alpha\right\} - \mathbf{1}\left\{\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) < \alpha\right\} \right| \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)$$

$$\overset{(ii)}{\leq} E\left(\mathbf{1}\left\{\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha\right| \leq \left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(G(\mathbf{Z}))\right|\right\} \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)$$

$$= P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha\right| \leq \left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(G(\mathbf{Z}))\right| \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)$$

$$\leq P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha\right| \leq c \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)$$

$$\quad + P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(G(\mathbf{Z}))\right| > c \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)$$

$$\overset{(iii)}{\leq} 2c + P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(G(\mathbf{Z}))\right| > c \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)$$

where (i) follows by (H.13), (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 (H.13). Now we take expectation on both sides, obtaining

$$E\left|P\left(\hat{F}(G(\mathbf{Z})) < \alpha \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha\right| \tag{H.14}$$

$$\leq 2c + P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(G(\mathbf{Z}))\right| > c\right)$$

$$\leq 2c + P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(\bar{G}(\mathbf{Z}))\right| > c\right) + P\left(\{\bar{Z}_t\}_{t \in A_R} \neq \{Z_t\}_{t \in A_R}\right)$$

$$\leq 2c + P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(\bar{G}(\mathbf{Z}))\right| > c\right) + \beta_{\text{mixing}}(k - T_* + 1)$$

$$\leq 2c + c^{-1} E\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(\bar{G}(\mathbf{Z}))\right| + \beta_{\text{mixing}}(k - T_* + 1)$$

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, $P(\mathcal{A}^c) \leq \gamma_{1,T} + \gamma_{2,T}$. Therefore,

$$E\left|\Psi\left(\bar{G}(\mathbf{Z}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right|$$

$$= E\left(\left|\Psi\left(\bar{G}(\mathbf{Z}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right| \times \mathbf{1}_{\mathcal{A}}\right)$$

$$\quad + E\left(\left|\Psi\left(\bar{G}(\mathbf{Z}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right| \times \mathbf{1}_{\mathcal{A}^c}\right)$$

$$\overset{(i)}{\leq} E\left(\left|\Psi\left(\bar{G}(\mathbf{Z}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right| \times \mathbf{1}_{\mathcal{A}}\right) + 2P\left(\mathcal{A}^c\right)$$

$$\leq \xi_T \varrho_T(|\widetilde{H}_R|) + 2P\left(\mathcal{A}^c\right) \leq \xi_T \varrho_T(m + 2k) + 2\gamma_{1,T} + 2\gamma_{2,T},$$

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

$$E\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(\bar{G}(\mathbf{Z}))\right|$$

$$\leq E\left|\Psi\left(\bar{G}(\mathbf{Z}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(\bar{G}(\mathbf{Z}))\right| + E\left|\Psi\left(\bar{G}(\mathbf{Z}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right|$$

$$\leq E\sup_{x \in \mathbb{R}}\left|\Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(x)\right| + \xi_T \varrho_T(m + 2k) + 2\gamma_{1,T} + 2\gamma_{2,T}$$

$$\overset{(i)}{\leq} C_4\left(\xi_T \varrho_T(m + 2k) + m^{1/2}\beta_{\text{mixing}}(m_1 - T_* + 1) + \sqrt{m_1/m} + \beta_{\text{mixing}}(k - T_* + 1) + \gamma_{1,T} + \gamma_{2,T}\right)$$

for a universal constant $C_4 > 0$, where (i) follows by (H.12).

Now we combine (H.14) 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

$$E\left|P\left(\hat{F}\left(G(\mathbf{Z})\right) < \alpha \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha\right|$$

$$\leq C_5\sqrt{\xi_T \varrho_T(m + 2k) + m^{1/2}\beta_{\text{mixing}}(m_1 - T_* + 1) + \sqrt{m_1/m} + \beta_{\text{mixing}}(k - T_* + 1) + \gamma_{1,T} + \gamma_{2,T}}.$$

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 $C_6 > 0$,

$$E\left|P\left(\hat{F}\left(G(\mathbf{Z})\right) < \alpha \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha\right|$$

$$\leq C_6\sqrt{\xi_T \varrho_T(m + 2k)} + C_6\left(m^{-1}(\log m)^{1/D_3}\right)^{1/4} + C_6\sqrt{\beta_{\text{mixing}}(k - T_* + 1)} + C_6\sqrt{\gamma_{1,T}} + C_6\sqrt{\gamma_{2,T}}.$$

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

H.2.1 Proof of Lemma H.6

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

$$\sup_{x \in \mathbb{R}} \left| \sqrt{T} L_T(x) - \sqrt{mK} \hat{F}(x) \right| \leq m - 1. \tag{H.15}$$

By Berbee's coupling (Lemma H.7), 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 (H.15) 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.

H.3 Proof of Theorem C.1

Recall $\mathbf{Z}^* = (Z_1^*, \ldots, Z_T^*)'$ with $Z_t^* = \left(Y_{1t}^N, Y_{2t}^N, \ldots, Y_{J+1t}^N, X_{1t}', \ldots, X_{J+1t}'\right)'$ for $1 \leq t \leq T_0 + T_*$. Let

$$\hat{p}_{\mathbf{Z}^*} = 1 - \hat{F}\left(S(\hat{u}(\mathbf{Z}^*))\right),$$

where $\hat{F}\left(x; \mathbf{Z}^*\right) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \mathbf{1}\left\{S\left(\hat{u}_\pi(\mathbf{Z}^*)\right) < x\right\}$ and $\hat{u}(\mathbf{Z}^*) = Y^N - \hat{P}^N$ with $\hat{P}^N$ computed using $\mathbf{Z}^*$. By the proof of Theorem 1, we have

$$|P\left(\hat{p}_{\mathbf{Z}^*} \leq \alpha\right) - \alpha| \leq C(\tilde{\delta}_T + \delta_T + \sqrt{\delta_T} + \gamma_T),$$

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

$$|P\left(\hat{p}_{\mathbf{Z}^*} > \alpha\right) - (1 - \alpha)| \leq C(\tilde{\delta}_T + \delta_T + \sqrt{\delta_T} + \gamma_T).$$

The desired result follows by observing that $\theta_t \in \mathcal{C}_{1-\alpha}(t)$ is the same event as $\hat{p}_{\mathbf{Z}^*} > \alpha$; this is simply because $\mathbf{Z}(\theta^0) = \mathbf{Z}^*$, where $\mathbf{Z}(\theta^0)$ is defined in Section 2.2.

H.4 Proof of Theorem D.1

Let $\{S^{(j)}(\hat{u})\}_{j=1}^n$ denoted the non-decreasing rearrangement of $\{S(\hat{u}_\pi) : \pi \in \Pi\}$, where $n = |\Pi|$, which we refer to as randomization quantiles. The p-value is

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

Note 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 proof proceeds in three steps. First, we show that exchangeability of the data implies exchangeability of the residuals. Second, we show that exchangeability of the residuals implies that $P(\hat{p} \leq \alpha) \leq \alpha$. Third, we show that if there are no ties, $\alpha - 1/n \leq P(\hat{p} \leq \alpha)$. The proof follows from standard arguments (e.g., Hoeffding, 1952; Romano, 1990; Chernozhukov et al., 2018; Lei et al., 2018).

Step 1: By the iid or exchangeability property of data, we have that

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

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

Step 2: Note that Π_{all} and $\Pi_\to$ form groups in the sense that $\Pi\pi = \Pi$ for all $\pi \in \Pi$. Therefore, the randomization quantiles are invariant,

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

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

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

Step 3: By continuity of the distribution of $\{S(\hat{u}_\pi)\}_{\pi \in \Pi}$, 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

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

The result now follows by similar arguments as in Step 2.

H.5 Proof of Lemma 1

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 $1 \le j \le J$. (2) there exist constants $c_1, c_2 > 0$ such that $E|X_{jt}u_t|^2 \ge c_1$ and $E|X_{jt}u_t|^3 \le c_2$ for any $1 \le j \le J$ and $1 \le t \le T$; (3) for each $1 \le j \le J$, the sequence $\{X_{jt}u_t\}_{t=1}^{T}$ is β-mixing and the β-mixing coefficient satisfies that $\beta(t) \le a_1 \exp(-a_2 t^\tau)$, where $a_1, a_2, \tau > 0$ are constants. (4) there exists a constant $c_3 > 0$ such that $\max_{1 \le j \le J} \sum_{t=1}^{T} X_{jt}^2 u_t^2 \le c_3^2 T$ 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 $(X_t'\delta)^2 \le \ell_T \|X\delta\|_2^2/T$, for all $w+\delta \in \mathcal{W}$ with probability $1 - o(1)$ for $T_0 + 1 \le t \le T$ 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 $1 - o(1)$

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

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

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

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

Lemma H.8. *Suppose that (1) $E(u_t X_{jt}) = 0$ for $1 \le j \le J$. (2) $\max_{1 \le j \le J, 1 \le t \le T} E|X_{jt}u_t|^3 \le K_1$ for a constant $K_1 > 0$. (3) $\min_{1 \le j \le J, 1 \le t \le T} E|X_{jt}u_t|^2 \ge K_2$ for a constant $K_2 > 0$. (4) For each $1 \le j \le J$, $\{X_{jt}u_t\}_{t=1}^{T}$ is β-mixing and the β-mixing coefficients satisfy $\beta(s) \le D_1 \exp\left(-D_2 s^\tau\right)$ 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 (Lemma H.7), 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) $P\left(\bigcup_{t\in H_i}\{\tilde{W}_{j,t} \neq W_{j,t}\}\right) \leq k\beta(m)$.

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 $0 \leq x \leq C_0 k^{1/6}$,

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

where $\Phi(\cdot)$ is the cdf of $N(0,1)$. Therefore, for any $0 \leq x \leq C_0 k^{1/6}$,

$$P\left(\left|\frac{\sum_{t\in H_i}W_{j,t}}{\sqrt{\sum_{t\in H_i}W_{j,t}^2}}\right| > x\right) \leq P\left(\left|\frac{\sum_{t\in H_i}\tilde{W}_{j,t}}{\sqrt{\sum_{t\in H_i}\tilde{W}_{j,t}^2}}\right| > x\right) + P\left(\bigcup_{t\in H_i}\{\tilde{W}_{j,t} \neq W_{j,t}\}\right)$$

$$\leq C_1\left(1 - \Phi(x)\right) + k\beta(m). \quad \text{(H.16)}$$

The Cauchy-Schwarz inequality implies

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

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

Hence,

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

It follows that for any $0 \leq x \leq C_0 k^{1/6}\sqrt{m}$,

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

$$= P\left(\sum_{i=1}^{m}\left(\frac{\sum_{t\in H_i}W_{j,t}}{\sqrt{\sum_{t\in H_i}W_{j,t}^2}}\right)^2 > x^2\right) \leq \sum_{i=1}^{m}P\left(\left|\frac{\sum_{t\in H_i}W_{j,t}}{\sqrt{\sum_{t\in H_i}W_{j,t}^2}}\right| > \frac{x}{\sqrt{m}}\right)$$

$$\stackrel{(i)}{\leq} m\left[C_1\left(1 - \Phi(x/\sqrt{m})\right) + k\beta(m)\right] \stackrel{(ii)}{\leq} C_1 m\sqrt{\frac{m}{2\pi}}x^{-1}\exp\left(-\frac{x^2}{2m}\right) + D_1 km\exp\left(-D_2 m^\tau\right)$$

$$< C_1 m^{3/2}x^{-1}\exp\left(-\frac{x^2}{2m}\right) + D_1 T\exp\left(-D_2 m^\tau\right)$$

where (i) follows by (H.16) 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 $0 \leq x \leq C_0 k^{1/6}\sqrt{m}$,

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

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 $x \ll C_0 k^{1/6}\sqrt{m}$ 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 (H.16) holds with C_1 large enough. The proof is complete. $\square$

Now we are ready to prove Lemma 1.

Proof of Lemma 1. 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

$$\|X\Delta\|_2^2 \leq 2u'X\Delta \stackrel{(i)}{\leq} 2\|Xu\|_\infty\|\Delta\|_1 \stackrel{(ii)}{\leq} 4K\|Xu\|_\infty, \tag{H.17}$$

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 H.8, 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 $P\left(\max_{1\leq j\leq J}\sum_{t=1}^T X_{jt}^2 u_t^2 \leq c_3^2 T\right) \to 1$, it follows that

$$P\left(\max_{1\leq j\leq J}\left|\sum_{t=1}^T X_{jt}u_t\right| > \kappa c_3[\log(T\vee J)]^{(1+\tau)/(2\tau)}\sqrt{T}\right) = o(1). \tag{H.18}$$

Part (1) follows by combining (H.17) and (H.18). Part (2) follows by part (1) and $\ell_T B_T = o(1)$. □

H.6   Proof of Lemma 2

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) $N^{-1} \sum_{i=1}^{N} E\|T^{-1/2} \sum_{t=1}^{T} F_t u_{it}\|_2^2 \leq D_0$.

(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, $N^{-1/2} \sum_{i=1}^{N} \lambda_i u_{it} \to^d N(0, \Gamma_t)$ 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, $T^{-1/2} \sum_{t=1}^{T} F_t u_{it} \to^d N(0, \Phi_i)$ as $T \to \infty$, where $\Phi_i = \lim_{T \to \infty} T^{-1} \sum_{s=1}^{T} \sum_{t=1}^{T} E(F_t F_s' u_{is}u_{it})$. (9) $N^{-1} \sum_{i=1}^{N} \lambda_i \lambda_i' \to \Sigma_\Lambda$ and $T^{-1} \sum_{t=1}^{T} F_t F_t' = \Sigma_F + o_P(1)$ 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):

$$\hat{\lambda}_1' \hat{F}_t - \lambda_1' F_t = \left(\hat{F}_t - H'F_t \right)' H^{-1} \lambda_1 + \hat{F}_t'(\hat{\lambda}_1 - H^{-1}\lambda_1). \tag{H.19}$$

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),

$$\sum_{t=1}^{T} \|\hat{F}_t - H'F_t\|_2^2 = O_P(T/\delta_{NT}^2), \tag{H.20}$$

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

$$\sum_{t=1}^{T} \|\hat{F}_t\|_2^2 = \sum_{t=1}^{T} \text{trace}(\hat{F}_t \hat{F}_t') = \text{trace}(\hat{F}'\hat{F}) = kT. \tag{H.21}$$

By Theorem 2 of Bai (2003),

$$\hat{\lambda}_1 = H^{-1}\lambda_1 + O_P(\max\{T^{-1/2}, N^{-1}\}). \tag{H.22}$$

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,

$$\|H^{-1}\lambda_1\|_2 = O_P(1). \tag{H.23}$$

Step 2: prove the desired result.

Therefore,

$$\sum_{t=1}^{T} \left(\hat{\lambda}_1'\hat{F}_t - \lambda_1'F_t\right)^2 \overset{(i)}{\leq} 2\sum_{t=1}^{T} \left[\left(\hat{F}_t - H'F_t\right)' H^{-1}\lambda_1\right]^2 + 2\sum_{t=1}^{T} \left[\hat{F}_t'(\hat{\lambda}_1 - H^{-1}\lambda_1)\right]^2$$

$$\leq 2\sum_{t=1}^{T} \|\hat{F}_t - H'F_t\|_2^2 \times \|H^{-1}\lambda_1\|_2^2 + 2\sum_{t=1}^{T} \|\hat{F}_t\|_2^2 \times \|\hat{\lambda}_1 - H^{-1}\lambda_1\|_2^2$$

$$\overset{(ii)}{=} O_P(T/\delta_{NT}^2) \times O_P(1) + 2kT \times O_P(\max\{T^{-1}, N^{-2}\})$$

$$= O_P(T/\delta_{NT}^2),$$

where (i) follows by (H.19) and the elementary inequality of $(a + b)^2 \leq 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$ and (ii) follows by (H.20), (H.21), (H.22) and (H.23). Since $n = |\Pi| = T$ for moving block permutation, we have

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

Finally, notice that Theorem 3 of Bai (2003) implies $\hat{\lambda}_1'\hat{F}_t - \lambda_1'F_t = O_P(1/\delta_{NT})$. The proof is complete.

H.7 Proof of Lemma 3

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

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

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),

$$\hat{\lambda}_1 = T^{-1}\hat{F}'(Y_1^N - X_1\hat{\beta}) = T^{-1}\hat{F}'(u_1 + F\lambda_1 - X_1\Delta_\beta). \tag{H.24}$$

Notice that

$$|\hat{u}_{1t} - u_{1t}|^2 = \left| F_t'\lambda_1 - \hat{F}_t'\hat{\lambda}_1 - X_{1t}'\Delta_\beta \right|^2$$

$$\overset{(i)}{=} \left| F_t'\lambda_1 - T^{-1}(H'F_t + \Delta_{F,t})'\hat{F}'(u_1 + F\lambda_1 - X_1\Delta_\beta) - X_{1,t}'\Delta_\beta \right|^2$$

$$\leq \left[\left| F_t'\left(I_k - H\hat{F}'F/T\right)\lambda_1 \right| + \left| T^{-1}\Delta_{F,t}'\hat{F}'F\lambda_1 \right| + \left| T^{-1}\hat{F}_t'\hat{F}'(u_1 - X_1\Delta_\beta) \right| + |X_{1t}'\Delta_\beta| \right]^2$$

$$\lesssim \left[F_t'\left(I_k - H\hat{F}'F/T\right)\lambda_1 \right]^2 + \left[T^{-1}\Delta_{F,t}'\hat{F}'F\lambda_1 \right]^2 + \left[T^{-1}\hat{F}_t'\hat{F}'(u_1 - X_1\Delta_\beta) \right]^2 + [X_{1t}'\Delta_\beta]^2,$$

$$\tag{H.25}$$

where (i) follows by (H.24) and $\hat{F}_t = H'F_t + \Delta_{F,t}$. Therefore,

$$
\sum_{t=1}^{T} (\hat{u}_{1t} - u_{1t})^2 \lesssim \sum_{t=1}^{T} \left[F_t' \left(I_k - H\hat{F}'F/T \right) \lambda_1 \right]^2 + \sum_{t=1}^{T} \left[T^{-1} \Delta_{F,t}' \hat{F}'F\lambda_1 \right]^2
$$

$$
+ \sum_{t=1}^{T} \left[T^{-1} \hat{F}_t' \hat{F}'(u_1 - X_1\Delta_\beta) \right]^2 + \sum_{t=1}^{T} [X_{1t}'\Delta_\beta]^2
$$

$$
\overset{(i)}{=} \lambda_1' \left(I_k - H\hat{F}'F/T \right)' (F'F) \left(I_k - H\hat{F}'F/T \right) \lambda_1
$$

$$
+ T^{-2} \left(\hat{F}'F\lambda_1 \right)' (\Delta_F'\Delta_F) \left(\hat{F}'F\lambda_1 \right) + T^{-1} \left\| \hat{F}'(u_1 - X_1\Delta_\beta) \right\|_2^2 + \| X_1\Delta_\beta \|_2^2
$$

$$
\overset{(ii)}{=} O_P \left(T\|\Delta_\beta\|_2^2 + T\delta_{NT}^{-4} \right) + O_P \left(T\|\Delta_\beta\|_2^2 + T\delta_{NT}^{-2} \right) + O_P \left(1 + T\delta_{NT}^{-4} + T\|\Delta_\beta\|_2^2 \right) + O_P(T\|\Delta_\beta\|_2^2)
$$

$$
= O_P \left(1 + T\|\Delta_\beta\|_2^2 + T\delta_{NT}^{-2} \right),
$$

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 H.9, 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 (H.25), we have

$$
|\hat{u}_{1t} - u_{1t}| \leq \left| F_t' \left(I_k - H\hat{F}'F/T \right) \lambda_1 \right| + \left| T^{-1} \Delta_{F,t}' \hat{F}'F\lambda_1 \right| + \left| T^{-1} \hat{F}_t' \hat{F}'(u_1 - X_1\Delta_\beta) \right| + |X_{1t}'\Delta_\beta|
$$

$$
\overset{(i)}{\leq} \|F_t\|_2 \cdot \|\lambda_1\|_2 \cdot O_P \left(\|\Delta_\beta\|_2 + \delta_{NT}^{-2} \right) + O_P \left(T\|\Delta_\beta\|_2 + T\delta_{NT}^{-2} \right) \cdot T^{-1} \|F\lambda_1\|_2
$$

$$
+ T^{-1} \|\hat{F}_t\|_2 \cdot O_P \left(\sqrt{T} + T\delta_{NT}^{-2} + T\|\Delta_\beta\|_2 \right) + \|X_{1t}\|_2 \cdot \|\Delta_\beta\|_2 \overset{(ii)}{\leq} O_P(T^{-1/2}),
$$

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 H.9), whereas (ii) follows by $\|\hat{F}_t\|_2 = O_P(1)$ (Lemma H.9), $\|X_{1t}\|_2 = O_P(1)$, $\|F_t\|_2 = O_P(1)$, $\lambda_1 = O(1)$, $\|\Delta_\beta\|_2 = O_P(T^{-1})$ and $\|F\lambda_1\|_2 = O_P(\sqrt{T})$.

Lemma H.9. *Suppose that the assumption of Lemma 3 holds. Let δ_{NT}, H, Δ_F and u_1 be defined as in the proof of Lemma 3. 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 $1 \leq t \leq T$.*

81

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

$$H = O_P(1) \quad \text{and} \quad H^{-1} = O_P(1). \tag{H.26}$$

Notice that

$$I_k - H\hat{F}'F/T \overset{(i)}{=} I_k - H(FH + \Delta_F)'F/T = I_k - (HH')(F'F/T) - H\Delta_F'F/T$$
$$\overset{(ii)}{=} O_P(\|\Delta_\beta\|_2) + O_P(\delta_{NT}^{-2}) - H\Delta_F'F/T$$
$$\overset{(iii)}{=} O_P(\|\Delta_\beta\|_2) + O_P(\delta_{NT}^{-2}), \tag{H.27}$$

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 (H.26) 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):

$$T^{-1}\Delta_F'\Delta_F = O_P(\|\Delta_\beta\|_2^2) + O_P(\delta_{NT}^{-2}). \tag{H.28}$$

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

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

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

$$\|F'u_1\|_2 = O_P(\sqrt{T}). \tag{H.29}$$

Notice that

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

$$\stackrel{\text{(ii)}}{=} O_P(\sqrt{T}) + \left(O_P(T^{1/2}\|\Delta_\beta\|_2) + O_P(T\delta_{NT}^{-2})\right) + O_P(T\|\Delta_\beta\|_2),$$

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 (H.26) and (H.29). 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

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

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

Proof of part (6). Notice that

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

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. $\qquad \square$

H.8 Proof of Lemma 4

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 4. 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

$$Y_{jt}^N = \text{trace}(Q_{jt}'M) + u_{jt} \qquad \text{for} \quad 1 \leq j \leq N, \ 1 \leq t \leq T. \tag{H.30}$$

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

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

and

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

Plugging (H.30) into the above inequality and rearranging terms, we obtain

$$\sum_{t=1}^{T} \sum_{j=1}^{N} \left(\text{trace}(Q_{jt}'\Delta)\right)^2 \leq 2 \sum_{t=1}^{T} \sum_{j=1}^{N} u_{jt}\text{trace}(Q_{jt}'\Delta) = 2\text{trace}\left(\left[\sum_{t=1}^{T} \sum_{j=1}^{N} u_{jt}Q_{jt}\right]' \Delta\right)$$

$$\overset{\text{(i)}}{=} 2\text{trace}(u'\Delta)$$

$$\overset{\text{(ii)}}{\leq} 2\|u\| \cdot \|\Delta\|_*$$

$$\overset{\text{(iii)}}{\leq} 4K\|u\|, \tag{H.31}$$

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 H.10. Note that the conditions of Lemma H.10 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 (H.31) imply that

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

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

Lemma H.10. *Suppose that the following conditions hold:*
(i) $\{u_j\}_{j=1}^N$ *is independent across j conditional on M.*
(ii) $\max_{1 \leq j \leq N} T^{-1} \sum_{t=1}^{T} E(|u_{jt}|^{2\kappa_1} \mid M) = O_P(1)$ *for some constant $\kappa_1 > 1$.*
(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 $|T^{-1} \sum_{t=1}^{T} a_t| \leq [T^{-1} \sum_{t=1}^{T} |a_t|^\kappa]^{1/\kappa}$ for any $\kappa > 1$ (due to Liapunov's inequality). It follows that $T^{-1} \sum_{t=1}^{T} u_{jt}^2 \leq [T^{-1} \sum_{t=1}^{T} |u_{jt}|^{2\kappa_1}]^{1/\kappa_1}$, which

means

$$\left(\sum_{t=1}^T u_{jt}^2\right)^{\kappa_1} \leq T^{\kappa_1-1}\sum_{t=1}^T |u_{jt}|^{2\kappa_1}. \tag{H.32}$$

Hence,

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

where (i) follows by Liapunov's inequality, (ii) follows by (H.32) and (iii) follows by the assumption that $\max_{1\leq j\leq N}T^{-1}\sum_{t=1}^T E(|u_{jt}|^{2\kappa_1}\mid M) = O_P(1)$. Therefore, it follows, by Theorem 5.48 and Remark 5.49 of Vershynin (2010), that

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

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

H.9 Proof of Lemma 5

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,

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

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 $T_0 + 1 \le t \le T$ (due to the stationarity property of u_t).

H.10 Proof of Lemma 6

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

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

and

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

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

H.11 Proof of Lemma 7

We first derive the following result that is useful in proving Lemma 7.

Lemma H.11. *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 iid 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) $(T-K)^{-1} \sum_{t=K+1}^{T} x_t u_t = o_P(1)$; (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 $(T-K)^{-1} \sum_{t=K+1}^{T} x_t x_t'$ 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 $\{u_s : s \le t\}$. First notice that $\{x_t u_t\}_{t=K+1}^T$ is a martingale difference sequence with respect to the filtration $\{\mathcal{F}_t\}$. Since ε_t is a stationary process, we have that $E\|x_t u_t\|^2 = \sum_{j=1}^K E(\varepsilon_{t-j}^2 u_t^2) = \sum_{j=1}^K E(\varepsilon_{t-j}^2) E(u_t^2)$ is uniformly 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 $1 \le j \le K$, $(T-K)^{-1} \sum_{t=K+1}^T \varepsilon_{t-j}^2 = E(\varepsilon_{t-j}^2) + o_P(1)$. 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

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

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

Now we are ready to prove Lemma 7.

Proof of Lemma 7. 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

$$\hat{\varepsilon}_t = \delta_t + \varepsilon_t = \delta_t + x_t'\rho + u_t = \delta_t + (\hat{x}_t - \Delta_t)'\rho + u_t = \hat{x}_t'\rho + \tilde{u}_t. \tag{H.33}$$

Therefore,

$$\hat{\rho} = \left(\sum_{t=K+1}^T \hat{x}_t \hat{x}_t'\right)^{-1} \left(\sum_{t=K+1}^T \hat{x}_t \hat{\varepsilon}_t\right) = \left(\sum_{t=K+1}^T \hat{x}_t \hat{x}_t'\right)^{-1} \left(\sum_{t=K+1}^T \hat{x}_t(\hat{x}_t'\rho + \tilde{u}_t)\right)$$

$$= \rho + \left(\sum_{t=K+1}^T \hat{x}_t \hat{x}_t'\right)^{-1} \left(\sum_{t=K+1}^T \hat{x}_t \tilde{u}_t\right). \tag{H.34}$$

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,

$$\sum_{t=K+1}^{T} \|\Delta_t\|^2 = \sum_{t=K+1}^{T} \sum_{s=t-1}^{t-K} \delta_s^2 \le K \sum_{t=1}^{T} \delta_t^2 \overset{(i)}{=} o_P(T), \tag{H.35}$$

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

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

$$\le 2 \sum_{t=K+1}^{T} \|x_t\| \cdot \|\Delta_t\| + \sum_{t=K+1}^{T} \|\Delta_t\|^2$$

$$\le 2 \sqrt{\left(\sum_{t=K+1}^{T} \|x_t\|^2 \right) \left(\sum_{t=K+1}^{T} \|\Delta_t\|^2 \right)} + \sum_{t=K+1}^{T} \|\Delta_t\|^2 \overset{(i)}{=} o_P(T), \tag{H.36}$$

where (i) follows by (H.35) and Lemma H.11. 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 H.11, 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

$$\left[(T-K)^{-1} \sum_{t=K+1}^{T} \hat{x}_t \hat{x}_t' \right]^{-1} = O_P(1). \tag{H.37}$$

Step 2: show that $(T-K)^{-1} \sum_{t=K+1}^{T} \hat{x}_t \tilde{u}_t = o_P(1)$.
By Lemma H.11, we have

$$(T-K)^{-1} \sum_{t=K+1}^{T} x_t u_t = o_P(1). \tag{H.38}$$

Notice that

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

$$\leq \frac{1}{T-K} \sum_{t=K+1}^{T} (\|\Delta_t u_t\| + \|x_t a_t\| + \|\Delta_t a_t\|)$$

$$\leq \sqrt{\left(\frac{1}{T-K} \sum_{t=K+1}^{T} \|\Delta_t\|^2\right) \left(\frac{1}{T-K} \sum_{t=K+1}^{T} u_t^2\right)}$$

$$+ \sqrt{\left(\frac{1}{T-K} \sum_{t=K+1}^{T} \|x_t\|^2\right) \left(\frac{1}{T-K} \sum_{t=K+1}^{T} a_t^2\right)}$$

$$+ \sqrt{\left(\frac{1}{T-K} \sum_{t=K+1}^{T} \|\Delta_t\|^2\right) \left(\frac{1}{T-K} \sum_{t=K+1}^{T} a_t^2\right)}. \quad \text{(H.39)}$$

We observe that

$$\sum_{t=K+1}^{T} a_t^2 = \sum_{t=K+1}^{T} (\delta_t - \Delta_t'\rho)^2 \leq 2 \sum_{t=K+1}^{T} \delta_t^2 + 2 \sum_{t=K+1}^{T} (\Delta_t'\rho)^2$$

$$\leq 2 \sum_{t=1}^{T} \delta_t^2 + 2\|\rho\|^2 \sum_{t=K+1}^{T} \|\Delta_t\|^2 \overset{\text{(i)}}{=} O_P(T), \quad \text{(H.40)}$$

where (i) follows by (H.35) and the assumption of $T^{-1}\sum_{t=1}^{T}\delta_t^2 = o_P(1)$. Combining (H.39) with (H.35) and (H.40), we obtain

$$\left\|\frac{1}{T-K} \sum_{t=K+1}^{T} (\hat{x}_t \tilde{u}_t - x_t u_t)\right\|$$

$$\leq \sqrt{o_P(1) \left(\frac{1}{T-K} \sum_{t=K+1}^{T} u_t^2\right)} + \sqrt{\left(\frac{1}{T-K} \sum_{t=K+1}^{T} \|x_t\|^2\right) o_P(1)} + \sqrt{o_P(1) \times o_P(1)} \overset{\text{(i)}}{=} o_P(1),$$

$$\text{(H.41)}$$

where (i) follows by Lemma H.11. Now we combine (H.38) and (H.41), obtaining

$$(T-K)^{-1} \sum_{t=K+1}^{T} \hat{x}_t \tilde{u}_t = o_P(1). \quad \text{(H.42)}$$

By (H.34) together with (H.37) and (H.42), it follows that

$$\hat{\rho} - \rho = o_P(1). \quad \text{(H.43)}$$

Step 3: show the desired result.

Recall that $\hat{u}_t = \hat{\varepsilon}_t - \hat{x}'_t \hat{\rho}$. Hence,

$$\hat{u}_t - u_t = (\hat{\varepsilon}_t - \hat{x}'_t \hat{\rho}) - u_t \overset{(i)}{=} (\hat{x}'_t(\rho - \hat{\rho}) + \tilde{u}_t) - u_t = \hat{x}'_t(\rho - \hat{\rho}) + a_t, \qquad \text{(H.44)}$$

where (i) follows by (H.33). Therefore, we have

$$\sum_{t=K+1}^{T} (\hat{u}_t - u_t)^2 = \sum_{t=K+1}^{T} (\hat{x}'_t(\rho - \hat{\rho}) + a_t)^2$$

$$\leq 2 \sum_{t=K+1}^{T} (\hat{x}'_t(\hat{\rho} - \rho))^2 + 2 \sum_{t=K+1}^{T} a_t^2$$

$$\leq 2\|\hat{\rho} - \rho\|^2 \sum_{t=K+1}^{T} \|\hat{x}_t\|^2 + 2 \sum_{t=K+1}^{T} a_t^2$$

$$= 2\|\hat{\rho} - \rho\|^2 \left(\sum_{t=K+1}^{T} \text{trace}(x_t x'_t) + \sum_{t=K+1}^{T} \text{trace}(\hat{x}_t \hat{x}'_t - x_t x'_t) \right) + 2 \sum_{t=K+1}^{T} a_t^2$$

$$\overset{(i)}{\leq} o_P(1) \times (O_P(T) + o_P(T)) + o_P(T) = o_P(T),$$

where (i) follows by (H.36), (H.43), (H.40) and Lemma H.11.

To see the pointwise result, we notice that by (H.44) and (H.43), 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. $\qquad \square$

H.12   Proof of Lemma E.1

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.

H.13 Proof of Lemma E.2

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

$$\|\xi_H - \xi\|_\infty \leq \|\hat{\mu}_H - \hat{\mu}\|_\infty + \|\hat{\Sigma}_H - \hat{\Sigma}\|_\infty \|\hat{\beta}\|_1 \leq c_T(1 + K). \tag{H.45}$$

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,

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

91

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

$$\lambda^2 \Delta' \hat{\Sigma} \Delta - 2\lambda \xi' \Delta \geq 0. \tag{H.46}$$

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

$$\begin{aligned}
0 \leq 2\lambda \xi'_H \Delta - \lambda^2 \Delta' \hat{\Sigma}_H \Delta &= 2\lambda \xi' \Delta - \lambda^2 \Delta' \hat{\Sigma} \Delta + 2\lambda(\xi_H - \xi)' \Delta + \lambda^2 \Delta'(\hat{\Sigma} - \hat{\Sigma}_H)\Delta \\
&\leq 2\lambda \xi' \Delta - \lambda^2 \Delta' \hat{\Sigma} \Delta + 2\lambda \|\xi_H - \xi\|_\infty \|\Delta\|_1 + \lambda^2 \|\hat{\Sigma} - \hat{\Sigma}_H\|_\infty \|\Delta\|_1^2.
\end{aligned}$$

It follows, by (H.45) and $\|\Delta\|_1 \leq \|\hat{\beta}_H\|_1 + \|\hat{\beta}\|_1 \leq 2K$, that

$$\lambda^2 \Delta' \Sigma \Delta - 2\lambda \xi' \Delta \leq 2\lambda \|\xi_H - \xi\|_\infty \|\Delta\|_1 + \lambda^2 \|\hat{\Sigma} - \hat{\Sigma}_H\|_\infty \|\Delta\|_1^2 \leq 4c_T(1+K)K + 4c_T K^2. \tag{H.47}$$

Since (H.46) and (H.47) hold for any $\lambda \in [0, 1]$, we have that

$$0 \leq \lambda^2 \Delta' \hat{\Sigma} \Delta - 2\lambda \xi' \Delta \leq 4c_T K(2K + 1) \qquad \forall \lambda \in [0, 1]. \tag{H.48}$$

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 (H.48), for any $\lambda \in [0, 1]$,

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

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

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

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

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

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 $P(\mathcal{M}) \geq 1 - \gamma_{1,T} - \gamma_{2,T} - \gamma_{3,T}$.

H.14 Proof of Lemma F.1

For notational simplicity, we write $\hat{\beta}$ instead of $\hat{\beta}_\lambda$ and $c_T = |H|$. Let $\hat{\Omega} = X'X$, $\tilde{\Omega} = \tilde{X}'\tilde{X}$, $\hat{\mu} = X'u$ and $\tilde{\mu} = \tilde{X}'\tilde{u}$. We work on the event on which $\kappa_1 T \leq \lambda_{\min}(\hat{\Omega}) \leq \lambda_{\max}(\hat{\Omega}) \leq \kappa_2 T$, $\|\hat{\mu} - \tilde{\mu}\|_2 \leq \kappa_3\sqrt{Jc_T}$ and $\|\hat{\Omega} - \tilde{\Omega}\| \leq \kappa_4(c_T + J)$. Let $q_1 \geq q_2 \geq \cdots \geq q_J$ denote the eigenvalues of $\hat{\Omega}$.

Step 1: show inconsistency.

Notice that

$$\begin{aligned}
\hat{\beta} &= (\hat{\Omega} + \lambda I)^{-1}X'(X\beta + u) \\
&= (\hat{\Omega} + \lambda I)^{-1}\hat{\Omega}\beta + (\hat{\Omega} + \lambda I)^{-1}X'u \\
&= \beta + \left[(\hat{\Omega} + \lambda I)^{-1}\hat{\Omega} - I\right]\beta + (\hat{\Omega} + \lambda I)^{-1}X'u \\
&= \beta + (\hat{\Omega} + \lambda I)^{-1}\left[\hat{\Omega} - (\hat{\Omega} + \lambda I)\right]\beta + (\hat{\Omega} + \lambda I)^{-1}X'u \\
&= \beta - \lambda(\hat{\Omega} + \lambda I)^{-1}\beta + (\hat{\Omega} + \lambda I)^{-1}X'u.
\end{aligned} \tag{H.49}$$

Therefore,

$$\begin{aligned}
&E\left(\|X(\hat{\beta} - \beta)\|_2^2 \mid X\right) \\
&= \|\lambda X(\hat{\Omega} + \lambda I)^{-1}\beta\|_2^2 + E\left(\|X(\hat{\Omega} + \lambda I)^{-1}X'u\|_2^2 \mid X\right) \\
&= \lambda^2\beta'(\hat{\Omega} + \lambda I)^{-1}\hat{\Omega}(\hat{\Omega} + \lambda I)^{-1}\beta + E\left(u'X(\hat{\Omega} + \lambda I)^{-1}\hat{\Omega}(\hat{\Omega} + \lambda I)^{-1}X'u \mid X\right)
\end{aligned}$$

93

$$= \lambda^2 \beta' (\hat{\Omega} + \lambda I)^{-1} \hat{\Omega} (\hat{\Omega} + \lambda I)^{-1} \beta + E \left(\text{trace} \left[(\hat{\Omega} + \lambda I)^{-1} \hat{\Omega} (\hat{\Omega} + \lambda I)^{-1} X' u u' X \right] \mid X \right)$$

$$= \lambda^2 \beta' (\hat{\Omega} + \lambda I)^{-1} \hat{\Omega} (\hat{\Omega} + \lambda I)^{-1} \beta + \text{trace} \left[(\hat{\Omega} + \lambda I)^{-1} \hat{\Omega} (\hat{\Omega} + \lambda I)^{-1} \hat{\Omega} \right].$$

Notice that

$$\text{trace} \left[(\hat{\Omega} + \lambda I)^{-1} \hat{\Omega} (\hat{\Omega} + \lambda I)^{-1} \hat{\Omega} \right] = \sum_{i=1}^{J} q_i^2 (q_i + \lambda)^{-2} \gtrsim \frac{T^2 J}{(T + \lambda)^2}$$

and

$$\lambda^2 \beta' (\hat{\Omega} + \lambda I)^{-1} \hat{\Omega} (\hat{\Omega} + \lambda I)^{-1} \beta \geq \lambda^2 \|\beta\|_2^2 \min_{1 \leq i \leq J} q_i (q_i + \lambda)^{-2} \gtrsim \frac{\lambda^2 T}{(T + \lambda)^2}.$$

Then we have

$$E \left[T^{-1} \|X(\hat{\beta} - \beta)\|_2^2 \mid X \right] \gtrsim \frac{TJ + \lambda^2}{(T + \lambda)^2} \gtrsim \frac{TJ}{(T + \lambda)^2} + \frac{\lambda^2}{(T + \lambda)^2}.$$

Since $\lambda \asymp T$, we have $E(\|X(\hat{\beta} - \beta)\|_2^2) \gtrsim T$.

Step 2: show stability.

Similar to (H.49), we notice that the perturbed estimator would be

$$\tilde{\beta} = \beta - \lambda (\tilde{\Omega} + \lambda I)^{-1} \beta + (\tilde{\Omega} + \lambda I)^{-1} \tilde{\mu}.$$

Then we bound

$$X_t'(\hat{\beta} - \tilde{\beta}) = \lambda X_t' \left[(\tilde{\Omega} + \lambda I)^{-1} - (\hat{\Omega} + \lambda I)^{-1} \right] \beta + X_t' \left[(\hat{\Omega} + \lambda I)^{-1} \hat{\mu} - (\tilde{\Omega} + \lambda I)^{-1} \tilde{\mu} \right]$$

$$= \lambda X_t' \left[(\tilde{\Omega} + \lambda I)^{-1} - (\hat{\Omega} + \lambda I)^{-1} \right] \beta + X_t' \left[(\hat{\Omega} + \lambda I)^{-1} (\hat{\mu} - \tilde{\mu}) + \left((\hat{\Omega} + \lambda I)^{-1} - (\tilde{\Omega} + \lambda I)^{-1} \right) \tilde{\mu} \right]$$

$$= X_t' \left[(\tilde{\Omega} + \lambda I)^{-1} - (\hat{\Omega} + \lambda I)^{-1} \right] (\lambda \beta - \tilde{\mu}) + X_t' (\hat{\Omega} + \lambda I)^{-1} (\hat{\mu} - \tilde{\mu}).$$

Now we notice that

$$\|(\tilde{\Omega} + \lambda I)^{-1} - (\hat{\Omega} + \lambda I)^{-1}\| = \left\| (\tilde{\Omega} + \lambda I)^{-1} \left[(\hat{\Omega} + \lambda I) - (\tilde{\Omega} + \lambda I) \right] (\hat{\Omega} + \lambda I)^{-1} \right\|$$

$$= \|(\tilde{\Omega} + \lambda I)^{-1} (\hat{\Omega} - \tilde{\Omega}) (\hat{\Omega} + \lambda I)^{-1}\| = O_P \left(\frac{(J + c_T)}{(T + \lambda)^2} \right)$$

and $\|\lambda \beta - \tilde{\mu}\|_2 \leq \lambda + \|\tilde{\mu}\|_2 = O_P(\lambda + \sqrt{JT})$. This means that

$$\left| X_t' \left[(\tilde{\Omega} + \lambda I)^{-1} - (\hat{\Omega} + \lambda I)^{-1} \right] (\lambda \beta - \tilde{\mu}) \right| = O_P \left(\frac{(J + c_T)}{(T + \lambda)^2} \left(\lambda + \sqrt{JT} \right) \sqrt{J} \right).$$

Moreover, we also have

$$\left| X_t'(\hat{\Omega} + \lambda I)^{-1}(\hat{\mu} - \tilde{\mu}) \right| \leq O_P\left(\sqrt{J} \frac{1}{\mu + \lambda} \sqrt{Jc_T} \right) = O_P\left(\frac{J\sqrt{c_T}}{T + \lambda} \right).$$

It follows that

$$\max_{1 \leq t \leq T} |X_t'(\hat{\beta} - \tilde{\beta})| = O_P\left(\frac{(J + c_T)}{(T + \lambda)^2} \left(\lambda + \sqrt{JT} \right) \sqrt{J} + \frac{J\sqrt{c_T}}{T + \lambda} \right).$$

Since $J \ll T$ and $T \gtrsim \lambda \gg J^{3/2}$, we have $\max_{1 \leq t \leq T} |X_t'(\hat{\beta} - \tilde{\beta})| = o_P(1)$.

I   Tables and Figures Appendix

Table I.1: Size Properties: $F_{2t} \overset{iid}{\sim} N(0,1)$, $\rho_\epsilon = \rho_u = 0$

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

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

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

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

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

Table I.2: Size Properties: $F_{2t} \overset{iid}{\sim} N(0,1)$, $\rho_\epsilon = \rho_u = 0.6$

	DGP1								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$
20	0.13	0.13	0.12	0.12	0.11	0.10	0.12	0.11	0.11
50	0.11	0.11	0.10	0.12	0.12	0.12	0.12	0.12	0.12
100	0.11	0.10	0.10	0.11	0.11	0.12	0.12	0.12	0.11
	DGP2								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$
20	0.12	0.12	0.11	0.12	0.11	0.11	0.12	0.11	0.11
50	0.11	0.11	0.11	0.11	0.12	0.12	0.12	0.11	0.12
100	0.10	0.10	0.11	0.11	0.11	0.12	0.11	0.11	0.12
	DGP3								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$
20	0.10	0.10	0.10	0.10	0.09	0.09	0.11	0.10	0.10
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	0.12	0.13	0.12
	DGP4								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$
20	0.11	0.11	0.10	0.11	0.10	0.10	0.12	0.11	0.12
50	0.11	0.11	0.11	0.11	0.11	0.11	0.12	0.12	0.12
100	0.10	0.10	0.11	0.11	0.11	0.11	0.11	0.11	0.11

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

Table I.3: Size Properties: $F_{2t} \sim N(t, 1)$, $\rho_\epsilon = \rho_u = 0$

|	DGP1								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$
20	0.10	0.10	0.10	0.08	0.09	0.09	0.07	0.08	0.09
50	0.11	0.09	0.10	0.09	0.09	0.09	0.09	0.08	0.09
100	0.10	0.10	0.10	0.10	0.09	0.09	0.10	0.10	0.09
	DGP2								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$
20	0.42	0.43	0.43	0.09	0.10	0.09	0.08	0.08	0.08
50	0.71	0.75	0.76	0.10	0.10	0.10	0.10	0.09	0.09
100	0.94	0.96	0.96	0.10	0.10	0.09	0.10	0.09	0.09
	DGP3								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$
20	0.45	0.43	0.44	0.51	0.48	0.49	0.08	0.09	0.08
50	0.78	0.75	0.78	0.82	0.80	0.80	0.09	0.09	0.09
100	0.97	0.97	0.97	0.98	0.97	0.98	0.09	0.10	0.10
	DGP4								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$
20	0.35	0.34	0.33	0.20	0.13	0.12	0.08	0.07	0.08
50	0.65	0.61	0.60	0.39	0.19	0.14	0.09	0.10	0.09
100	0.89	0.86	0.86	0.61	0.33	0.21	0.09	0.10	0.10

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

Table I.4: Size Properties: $F_{2t} \sim N(t,1)$, $\rho_\epsilon = \rho_u = 0.6$

	DGP1								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$
20	0.12	0.12	0.12	0.10	0.09	0.09	0.10	0.09	0.08
50	0.10	0.12	0.11	0.11	0.11	0.11	0.12	0.12	0.11
100	0.10	0.11	0.10	0.11	0.12	0.11	0.11	0.12	0.12

	DGP2								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$
20	0.44	0.47	0.49	0.12	0.12	0.13	0.10	0.10	0.10
50	0.74	0.77	0.77	0.11	0.11	0.11	0.12	0.12	0.12
100	0.95	0.96	0.96	0.11	0.11	0.11	0.11	0.11	0.11

	DGP3								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$
20	0.45	0.45	0.45	0.53	0.51	0.53	0.10	0.09	0.09
50	0.78	0.78	0.78	0.83	0.82	0.81	0.12	0.12	0.12
100	0.97	0.97	0.97	0.98	0.97	0.97	0.12	0.11	0.11

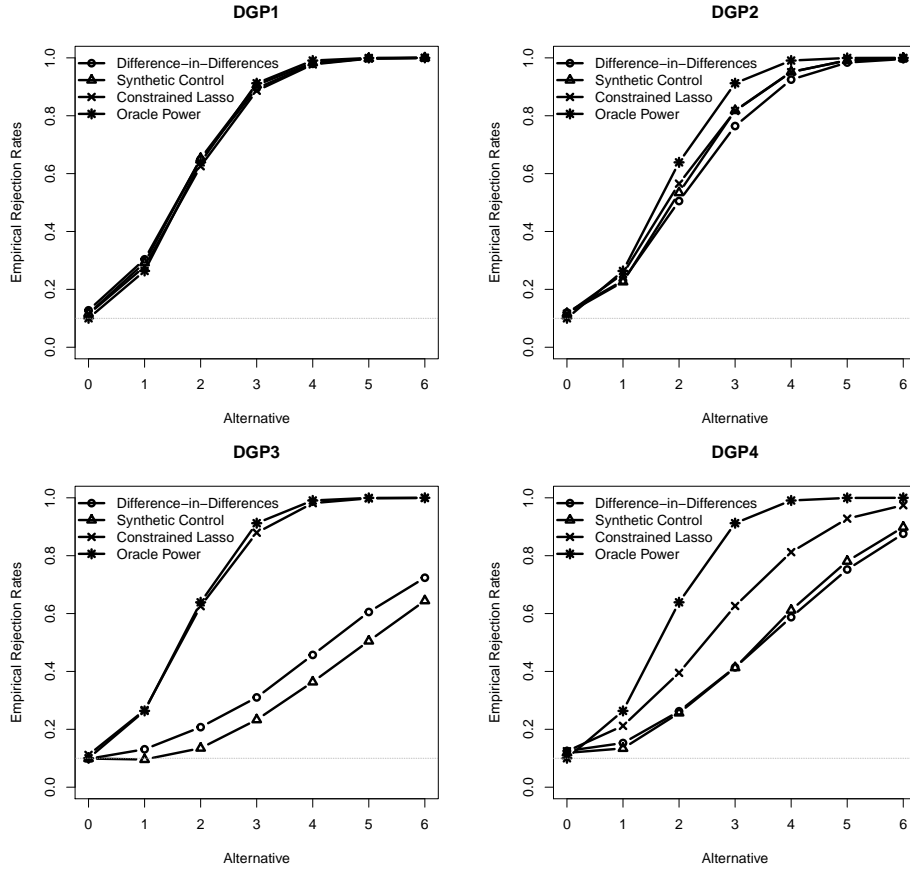|	DGP4								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$
20	0.38	0.37	0.36	0.21	0.14	0.12	0.10	0.10	0.10
50	0.66	0.63	0.63	0.39	0.20	0.14	0.12	0.12	0.11
100	0.89	0.86	0.86	0.62	0.34	0.20	0.10	0.11	0.11

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

Figure I.1: Finite Sample Size Properties ($\alpha = 0.1$)
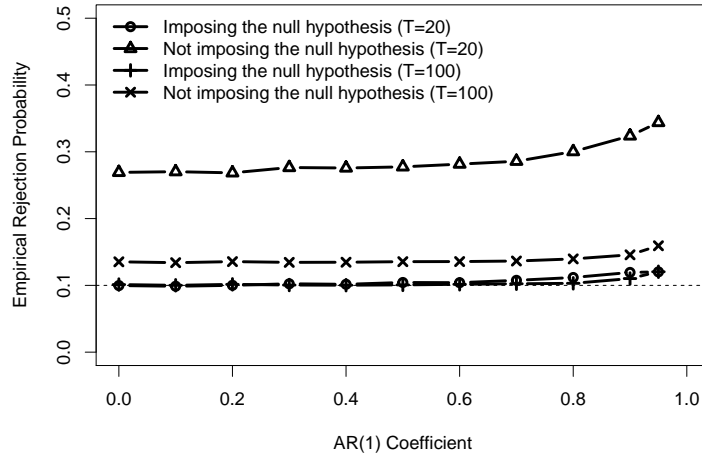
Notes: Empirical rejection probability from testing $H_0 : \theta_T = 0$. The data are generated as $Y_{1t}^N = \sum_{j=2}^{J+1} w_j Y_{jt}^N + u_t$, where $Y_{jt}^N \sim N(0,1)$ is iid across (j,t), $\{u_t\}$ is a Gaussian AR(1) process, $(w_2, \ldots, w_{J+1})' = (1/3, 1/3, 1/3, 0, \ldots, 0)'$, and $J = 50$. The weights are estimated using the canonical SC method (cf. Section 2.3.2).

Figure I.2: Power Curves

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