Chernozhukov, Victor; Wüthrich, Kaspar; Zhu, Yu

Working Paper

An exact and robust conformal inference method for counterfactual and synthetic controls

cemmap working paper, No. CWP62/17

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Chernozhukov, Victor; Wüthrich, Kaspar; Zhu, Yu (2017) : An exact and robust conformal inference method for counterfactual and synthetic controls, cemmap working paper, No. CWP62/17, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2017.6217

This Version is available at:
https://hdl.handle.net/10419/189808

cemmap
centre for microdata methods and practice

An exact and robust conformal inference method for counterfactual and synthetic controls

Victor Chernozhukov
Kaspar Wüthrich
Yu Zhu

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP62/17

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

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

V. Chernozhukov† K. Wüthrich‡ Y. Zhu§

December 26, 2017

Abstract

This paper introduces new inference methods for counterfactual and synthetic control methods for evaluating policy effects. Our inference methods work in conjunction with many modern and classical methods for estimating the counterfactual mean outcome in the absence of a policy intervention. Specifically, our methods work together with the difference-in-difference, canonical synthetic control, constrained and penalized regression methods for synthetic control, factor/matrix completion models for panel data, interactive fixed effects panel models, time series models, as well as fused time series panel data models. The proposed method has a double justification. (i) If the residuals from estimating the counterfactuals are exchangeable as implied, for example, by i.i.d. data, our procedure achieves exact finite sample size control without any assumption on the specific approach used to estimate the counterfactuals. (ii) If the data exhibit dynamics and serial dependence, our inference procedure achieves approximate uniform size control under weak and easy-to-verify conditions on the method used to estimate the counterfactual. We verify these condition for representative methods from each group listed above. Simulation experiments demonstrate the usefulness of our approach in finite samples. We apply our method to re-evaluate the causal effect of election day registration (EDR) laws on voter turnout in the United States.

1 Introduction

We consider the problem of making inference on the causal effect of a policy intervention in an aggregate time series setup with a single treated unit. The treated unit is observed for a number of periods before and after the intervention occurs. Often, there is additional information in the form of possibly very many untreated units, which can serve as controls. Such setups frequently arise in applied economic research and there are various different approaches to estimate the policy effects of interest. A non-exhaustive list of methods includes difference-in-differences methods (e.g., Ashenfelter and Card, 1985; Card and Krueger, 1994; Bertrand et al., 2004; Athey and Imbens, 2006; Angrist and Pischke, 2008), synthetic control models (e.g., Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015; Li, 2017), penalized regression models for synthetic controls (e.g., Valero, 2015; Doudchenko and Imbens, 2016), factor, matrix completion and interactive fixed effects models

*We are grateful to seminar participations at the University of Chicago, the University of Wisconsin Madison, UC Los Angeles, and UC San Diego for valuable comments. All errors are our own.

†email: vchern@mit.edu

‡email: kwuthrich@ucsd.edu

§email: yzhu6@uoregon.edu

for panel data (e.g., Bai, 2003; Pesaran, 2006; Bai, 2009; Kim and Oka, 2014; Gobillon and Magnac, 2016; Xu, 2017; Athey et al., 2017; Amjad et al., 2017), matching methods (e.g., Heckman et al., 1997, 1998; Dehejia and Wahba, 2002), as well as standard time series models. Doudchenko and Imbens (2016) and Gobillon and Magnac (2016) provide comparative overviews. We refer to these approaches as counterfactual and synthetic control methods (CSC) methods.

The main objective and contribution of this paper is to provide inference procedures for policy effects estimated by CSC methods. There are several practical issues which render inference in typical CSC setups challenging. First, the number of pre-treatment periods T_0 and, in particular, the number of post-treatment periods T_* are both small. Second, the data exhibit dynamics and serial dependence. Third, the number of (potential) control units J is of the same order as T_0. This leads to a need for some regularization. Finally, since there is only one treated unit and T_0 and T_* are small, treatment effects cannot be estimated consistently. This paper develops an inference approach to address these challenges.

We analyze a general counterfactual modeling framework (CMF) that nests and generalizes many traditional and new methods for counterfactual analysis. Specifically, we focus on models which are able to generate a mean-unbiased proxy P_t^N for the counterfactual outcome of the treated unit in the absence of the policy Y_{1t}^N:

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

The policy effect in period t is given by $\alpha_t = Y_{1t}^I - Y_{1t}^N$, where Y_{1t}^I is the counterfactual outcome of the treated unit with the policy. We are interested in testing hypotheses about the trajectory of policy effects in the post-intervention period: $\alpha = \{\alpha_t\}_{t=T_0+1}^{T_0+T_*}$. Specifically, we postulate a null trajectory $\alpha^o = \{\alpha_t^o\}_{t=T_0}^{T_0+T_*}$ and test the sharp null $H_0 : \alpha = \alpha^o$. We also consider testing hypotheses about a single time periods, $H_0 : \alpha_t = \alpha_t^o$, as well as in constructing pointwise confidence intervals for α_t. The basic idea of our testing procedure is the following. Under the sharp null hypothesis, we can construct Y_{1t}^N for each $t \in \{1, \ldots, T_0 + T_*\}$, estimate P_t^N, and back out the residuals $\hat{u}_t$. Under our assumptions, the distribution of $\{u_t\}$ in the post treatment period should be the same as that of $\{u_t\}$ in the pre-treatment period. We operationalize this idea by proposing a conformal/permutation inference procedure in which p-values are obtained by permuting the estimated residuals across the time series dimension. The proposed procedure has a double justification:[1]

(i) Exact Validity under Strong Assumptions.

If the residuals $\{\hat{u}_t\}$ are exchangeable, our inference procedure achieves finite sample (non-asymptotic) size control without any assumption on the method used to estimate P_t^N. Exchangeability of $\{\hat{u}_t\}$ is implied, for example, if the data are i.i.d. across time under the null, but holds more generally.

(ii) Approximate Validity under Weak Assumptions.

If the data exhibit dynamics and serial dependence, our inference procedure has an approximate finite sample justification under stationarity and weak dependence of u_t and easy-to-verify conditions (pointwise consistency and consistency in prediction norm) on the method used to estimate the counterfactual mean proxy P_t^N. These conditions can be verified for many different CSC methods. We provide concrete sets of sufficient conditions for a representative set of methods, including canonical synthetic control estimators, factor/matrix

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

completion models, interactive fixed effects estimators, Lasso, simple time series models, as well as fused time series panel data models.

To derive the theoretical properties of our inference procedure, we develop general results on exact and approximate permutation inference, which may be of independent interest in many conformal and permutation inference problems.

As part of developing our results, we introduce the ℓ_1-constrained least squares estimator (constrained Lasso) (e.g., Raskutti et al., 2011) as an essentially tuning free alternative to existing penalized regression estimators in settings with potentially many control units. Our analysis of constrained Lasso further provides new results for classical synthetic control estimators in settings with potentially very many control units.

We discuss two extensions of our main results. First, we show that our methods can be modified to test hypotheses about average effects $\bar{\alpha} = 1/T_* \sum_{t=T_0+1}^{T_0+T_*} \alpha_t$. Second, we propose easy-to-implement specification tests that allow us to assess the plausibility of the key assumptions underlying the proposed inference procedure.

Simulation experiments demonstrate favorable finite sample properties of the proposed inference procedures. To illustrate the practical usefulness of our methods, we revisit the analysis of the effect of Election Day Registration (EDR) laws on voter turnout in the United States by Xu (2017).

1.1 Related Literature

Conceptually, our procedure builds on the literature on conformal prediction (Vovk et al., 2005, 2009; Lei et al., 2013, 2017) and, more broadly, on the literature on permutation tests (Romano, 1990; Lehmann and Romano, 2005), which was started by Fisher (1935) in the context of randomization; see also Rubin (1984) for a Bayesian justification. Conformal inference, a form of permutation inference, is a distribution-free approach for forming prediction intervals. The basic idea is classical: Let $\{Y_1, \ldots, Y_T\}$ be a random sample drawn from a distribution P. To decide whether a new draw $Y_{T+1} = y$ should be included in the prediction set, we test the hypothesis that $Y_{T+1} = y$. A distribution-free and valid p-value can be constructed based on the quantile of the empirical distribution of the augmented sample $\{Y_1, \ldots, Y_n, y\}$. We still prefer to use the name "conformal" to designate a more specialized area of permutation inference that specializes on building predictive confidence intervals. Our analysis will deviate from the basic analysis of permutation inference, when we have to deal with *dependent data* and the fact that the models will be estimated. Our asymptotic results will be of independent interest for any type of permutation inference carried out for dependent data (see our Propositions 1 and 2). On a more general conceptual level, our approach is also connected to transformation-based approaches to model-free prediction (Politis, 2015).

The proposed inference procedure is further related to Andrews (2003)'s end-of-sample stability test based on subsampling. Besides a different focus (inference on policy effects vs. testing for structural breaks), there are several major differences. First, our procedure has exact validity under exchangeability and we obtain approximate finite sample bounds under weak conditions on the estimators, while such properties have not been established for Andrews (2003)'s test. The second major difference is that our test only requires stationarity and weak dependence of the stochastic process $\{u_t\}$, whereas Andrews (2003)'s test is based on stationarity of the data.[2] A third major difference is that our procedure works in conjunction with many modern high-dimensional esti-

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

mators, whereas Andrews (2003) focuses on low-dimensional GMM-type models. Hahn and Shi (2016, Section 5) informally suggest applying the end-of-sample stability test in the context of synthetic control methods and Ferman and Pinto (2017a) use a version of this test in the context of difference-in-differences approaches with few treated groups.

Our paper contributes to the literature on inference for CSC methods with few treated units. One part of the literature considers a finite population approach, which relies on the assumption that potential outcomes are fixed but a priori unknown and that, conditional on observables, the treatment assignment is random (Firpo and Possebom, 2017). These assumptions justify the application of permutation tests similar to Fisher (1935)'s randomization test. For instance, Abadie et al. (2010, 2015) permute which unit is assigned to the treatment and then compare the actual treatment effect estimates to the permutation distribution.[3] Firpo and Possebom (2017) and Ferman and Pinto (2017b) provide a comprehensive discussion of the theoretical aspects of such testing procedures. While finite population permutation approaches have traditionally been employed in conjunction with synthetic control methods, they can also be applied to a broader class of methods including difference-in-differences approaches, elastic net, and best subset selection, see, e.g., Doudchenko and Imbens (2016). Our approach will instead carry out the permutations over stochastic errors in the potential outcomes with respect to time, and not the cross-sectional units. These types of permutations rely on weak dependence of stochastic errors over time rather than exchangeability of the errors across treated units. While are results are for permutations across the time series dimension, our general results on exact and approximate permutation inference (Propositions 1 and 2) also apply to permutations across units (subject to switching indices). This provides a rigorous formal justification for the inference procedure of Abadie et al. (2010, 2015) under a set of sufficient conditions, which differ substantially from existing ones (e.g., Firpo and Possebom, 2017). We will justify our approach for a great variety of models to build counterfactual proxies for outcomes in the absence of the policy intervention, including many popular synthetic control, panel data, and fused time series panel models.

Another part of the literature considers asymptotic inference for CSC models. Asymptotic approaches often focus on testing hypotheses about average effect over time, $\bar{\alpha}$, and require that T_0 and often also T_* tend to infinity. Carvalho et al. (2017) derive the asymptotic distribution of $\bar{\alpha}$ in setups where the counterfactual is estimated based on Lasso and Li (2017) studies inference based on the constrained least squares estimator of Abadie et al. (2010, 2015). Xu (2017) proposes an asymptotic bootstrap inference procedure based on factor models, but leaves the formal justification of this procedure for future research. By contrast, our approach will instead be based on permutation distributions, and will be shown to be formally valid exactly under strong exchangeability assumptions and approximately valid under stationarity and weak dependence of $\{u_t\}$ and very weak conditions on the estimator for P_t^N. We verify these conditions for many different methods including constrained least squares estimators, Lasso, and factor models.

1.2 Plan of the Paper

The remainder of this paper is structured as follows. Section 2 introduces our basic modeling framework, the proposed inference method, and various models for counterfactuals P_t^N. In Section 3, we establish the finite sample validity of our procedure if $\{\hat{u}_t\}$ is exchangeable and the approximate

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

finite sample validity under stationarity and weak dependence of $\{u_t\}$ and weak and easy-to-verify high-level conditions on the estimator of P_t^N. Section 4 discusses two extensions of our procedure. In Section 5, we verify the high-level conditions for several representative CSC estimators. Section 6 presents simulation evidence on the finite sample properties of our estimators. In Section 7, we illustrate our procedure by reanalyzing the impact of EDR laws on voter turnout. Section 8 concludes. The appendix contains all proofs as well as some general results on the exact and approximate validity of conformal and permutation inference procedures which may be of independent interest.

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

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

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

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

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

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

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

Assumption 1 introduces potential outcomes, but also postulates an identifying assumption in the form of the existence of mean-unbiased proxies P_t^N such that

$$E\left(P_t^N\right) = E\left(Y_t^N\right), \quad t = 1, \ldots, T.$$

We will discuss specific panel data and time series models that postulate (and identify) what P_t^N is under a variety of conditions. Additional assumptions on the stochastic shock process $\{u_t\}$ will be introduced later, in essence requiring $\{u_t\}$ to be either i.i.d. or more generally a stationary and weakly dependent process. In principle, the treatment effect sequence $\{\alpha_t\}$ can be allowed to be random, and we can interpret our model and the results as holding conditional on a given $\{\alpha_t\}$. Hence, there is not much loss of generality in assuming that the sequence is fixed. Assumption 1 also postulates that the stochastic shock sequence will be invariant under the intervention. This is the key identifying assumption. In principle, we can relax this assumption by specifying for

example the scale and quantile shifts in the stochastic shocks that result from the policy, and then working with the resulting model; we leave this extension to future work. The CMF nests many traditional and new methods for counterfactual policy analysis, including difference-in-differences methods, canonical synthetic control, constrained and penalized regressions for synthetic control, factor/matrix completion models for panel data, interactive fixed effects panel models, univariate time series models, as well as fused time series panel data models.

Often, we have additional information in the form of untreated units, which can serve as controls. Specifically, suppose that there are $J \geq 1$ control units, where control units are indexed by $j = 2, \ldots, J+1$. We observe all units for all T periods, although this assumption can be relaxed. Let Y_{jt} denote the observed outcome for these untreated units. This observed outcome is equal to the outcome in the absence of the policy intervention, Y_{jt}^N, so that

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

For each unit, we may also observe a vector of covariates X_{jt}. This motivates a variety of strategies for modeling and identifying P_t^N as discussed below.

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

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

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

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

We are interested in testing hypotheses about $\alpha = (\alpha_{T_0+1}, \ldots, \alpha_T)'$. Our main hypothesis of interest is

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

where $\alpha^o = \left(\alpha_{T_0+1}^o, \ldots, \alpha_T^o\right)'$ is a postulated policy effect trajectory. Hypothesis (1) is a sharp null hypothesis. It fully determines the value of the counterfactual outcome with the treatment in the post treatment period since $Y_{1t}^N = Y_{1t}^I - \alpha_t = Y_{1t} - \alpha_t$. Our procedure can be extended to testing hypotheses about average effects as discussed in Section 4.1. While α^o can generally be an unrestricted function of t, it is sometimes useful and interesting to consider parametric hypotheses such as

$$\alpha_t^o = a_1^o + a_2^o(t - T_0), \quad t > T_0.$$

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

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

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

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

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

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

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

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

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

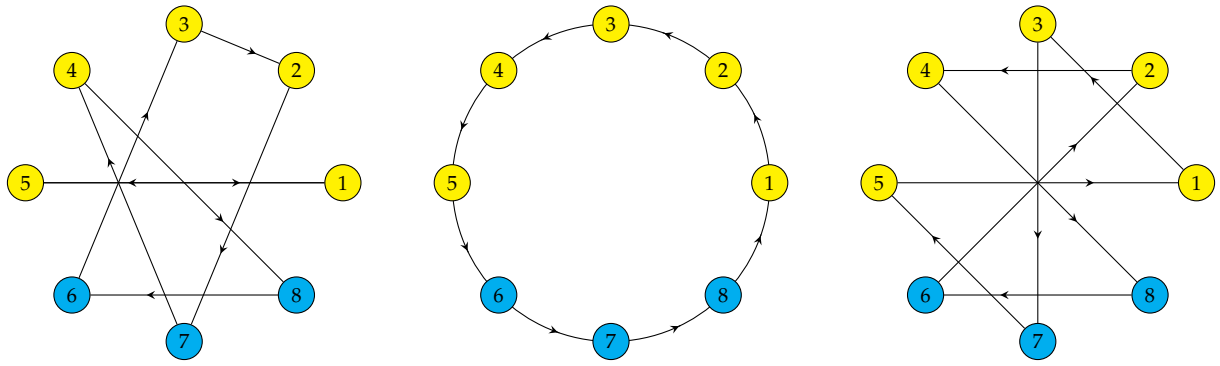
We use permutations to compute p-values. A permutation π is a one-to-one mapping $\pi : \{1, ..., T\} \mapsto \{1, ..., T\}$. We denote the set of all permutations under study as Π. Throughout the paper we assume that Π contains the identify map $\mathbb{I}$. We mainly focus on two different sets of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

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

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

Here we introduce other permutation groups, which we call the "i.i.d. block" and "overlapping block" permutations. To define the first group, we divide the data up into non-overlapping $K = T/m$ blocks of size m. Then we construct the "i.i.d" permutations of all blocks. Specifically, let $\{b_1, \ldots, b_K\}$ be the partition of $\{1, \ldots, T\}$, then we collect all the permutations π of these blocks, forming the "i.i.d. m-block" permutation Π_{mb}. Finally, by taking the composition $\Pi_{ob} = \Pi_{mb}\Pi_{\rightarrow}$ we create the "overlapping m block" group Π_{ob}, the permutation analog of the "overlapping block" bootstrap. These ideas are very close to bootstrap and/or subsampling, with the difference that our method will actually be exact under i.i.d. data and approximately valid for general data, with no limit distributions required. In our context choosing $m = T_*$ is natural, though other choices should work as well, similarly to the choice of block size in the time-series bootstrap.

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



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

For each $\pi \in \Pi$, let $\hat{u}_\pi = (\hat{u}_{\pi(1)}, \ldots, \hat{u}_{\pi(T)})'$ denote the vector of permuted residuals. We note that if the estimator used in approximating P_t^N is invariant to permutations of the data $\{Z_t\}$ across the time series dimension (which is the case for most of the estimators we consider in Sections 2.3 and 2.4), permuting $\{\hat{u}_t\}$ is equivalent to permuting $\{Z_t\}$.

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

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

where

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

An important special case of the testing problem (1) occurs if $T_* = 1$ in which case

$$H_0 : \alpha_T = \alpha_T^o \tag{2}$$

Finally, suppose we are interested in testing

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

for some fixed $t \geq T_0$. Hypothesis (3) can be tested by redefining Z as $\tilde{Z} = (Z_1, \ldots, Z_{T_0}, Z_t)'$ and testing (2). Pointwise confidence intervals for $t \in \{T_0 + 1, \ldots, T\}$ can be constructed by test inversion.

Next, we develop several models for generating the counterfactual proxies P_t^N.

2.3 Models for Counterfactual Proxies P_t^N via Synthetic Control and Panel Data

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

2.3.1 Difference-in-Differences Methods

The difference-in-difference model postulates

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

using an average of $J \geq 1$ of outcomes of control units as a proxy. This model automatically embeds the identifying information. The counterfactual can be estimated as

$$\hat{P}_t^N = \frac{1}{T_0}\sum_{t=1}^{T_0}\left(Y_{1t}^N - \frac{1}{J}\sum_{j=2}^{J+1} Y_{jt}\right) + \frac{1}{J}\sum_{j=2}^{J+1} Y_{jt}.$$

2.3.2 Synthetic Control and Constrained Least Squares Estimators

The canonical synthetic control postulates the model

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

where $w = (w_2, \ldots, w_{J+1})'$ is a vector of weights satisfying $w \geq 0$ and $\sum_{j=2}^{J} w_j = 1$.

We also need to impose an identification condition that allows us to identify the weights, for example[4]:

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

$$E\left(u_t\{Y_{jt}\}_{j=2}^{J+1}\right) = 0, \tag{4}$$

The counterfactual is estimated as

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

We focus on the following canonical SC estimator for w:[5]

$$\hat{w} = \arg\min_{w} \sum_{t=1}^{T_0}\left(Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt}\right)^2 \quad \text{s.t. } w \geq 0 \text{ and } \sum_{j=2}^{J} w_j = 1. \tag{5}$$

As an alternative, we can consider the more flexible model

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

[4]More generally, other exclusion restrictions could be used.

[5]We focus on the canonical problem (5) for concreteness. Abadie et al. (2010, 2015) consider a more generalized version, which also includes covariates into the estimation of the weights w. Doudchenko and Imbens (2016) refer to the estimator (5) as constrained regression.

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

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

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

$$(\hat{\mu}, \hat{w}) = \arg\min_{(\mu,w)} \sum_{t=1}^{T_0} \left(Y_{1t} - \mu - \sum_{j=2}^{J+1} w_j Y_{jt} \right)^2 \quad \text{s.t. } ||w||_1 \leq 1 \tag{6}$$

The advantage over other penalized regression methods discussed next is that constrained Lasso is essentially tuning free, and will be shown to be valid under very weak conditions. We will verify that these estimators are valid in our framework under weak conditions in setups with potentially many controls J. Finally, we note that it is straightforward to incorporate (transformations of) covariates X_{jt} into the estimation problems (5) and (6).

2.3.3 Penalized Regression Methods

Consider the following regression model for P_t^N

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

where μ is an intercept term $w = (w_2, \ldots, w_{J+1})'$ is a vector of weights. Here, we maintain the identifying assumption (SC). Under this assumption the counterfactual is estimated by

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

where

$$(\hat{\mu}, \hat{w}) = \arg\min_{(\mu,w)} \sum_{t=1}^{T_0} \left(Y_{1t} - \mu - \sum_{j=2}^{J+1} w_j Y_{jt} \right)^2 + \mathcal{P}(w) \tag{7}$$

where $\mathcal{P}(w)$ is a penalty function, which penalizes deviations away from zero. If it is desired to penalizes deviations away from other focal points w^0, for example, $w^0 = (1/J, ..., 1/J)$ 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. For example:

- Lasso (Tibshirani, 1996): $\mathcal{P}(w) = \lambda||w||_1$ where λ is a tuning parameter. A version is the Post-Lasso estimator, which refits the weights after removing variables with zero weight.
- Elastic Net (Zou and Hastie, 2005): $\mathcal{P}(w) = \lambda\left((1-\alpha)||w||_2 + \alpha||w_1||_1\right)$ where λ and α are tuning parameters.

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

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 sufficient set of conditions that exist in the literature.

2.3.4 Interactive Fixed Effects Models/Matrix Completion Models

Consider the following interactive fixed effects (FE) model for treated and untreated units:

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

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

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

The model leads to the following proxy:

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

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 (which allow for unbalanced data) applied to the model (8); see e.g. Bai (2009) and Hansen and Liao (2016).

Model (8) nests the classical factor model

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

which in turn covers the traditional linear FE model, in which

$$\lambda_i' F_t = \lambda_i + F_t.$$

There is a large body of work on these type of models; in econometrics these models are called interactive effects and augmented factor models and in statistics and machine learning they are called low-rank approximations and estimated through penalization methods or through universal singular value thresholding (upon imputing the missing entries with some reasonable proxies); see, e.g., Amjad et al. (2017) and Athey et al. (2017), where such methods are used for predicting counterfactual response, albeit they do not provide inference methods. Our proposal delivers a way to perform valid inference for policy effects in these models, including the recent new methods, even though we shall be focusing on Bai (2009)'s alternating least squares estimator when verifying our conditions. The results in Hansen and Liao (2016) imply that our high-level conditions hold for their estimator.

2.4 Models for Counterfactual Proxies via Time Series and Fused Models

2.4.1 Simple Time Series Models

If no control units are available, one can use time series models for the single unit exposed to the treatment. For example, consider the following autoregressive model:[6]

$$\begin{array}{l} Y_{1t}^N - \mu = \rho(Y_{1(t-1)}^N - \mu) + u_t \\ Y_{1t}^I - \mu = \rho(Y_{1(t-1)}^N - \mu) + \alpha_t + u_t \end{array} \quad \Big| \quad E(u_t) = 0, \quad \{u_t\} \text{ i.i.d.}, \quad t = 1, \ldots, T. \tag{10}$$

In this model the mean unbiased proxy is given by:

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

Note that the policy effect here is transitory, namely it does not feed-forward itself on the future values of Y_{1t}^I beyond the current values.[7] Under the null hypothesis, we can impute the unobserved counterfactual as $Y_{1t}^N = Y_{1t} - \alpha_t$, for $t > T_0$, and estimate the model using traditional time-series methods and we can make conformal inference based on the residuals.

The simplest form of the autoregressive model is the AR(K) process, where the $\rho(\cdot)$ take the form:

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

where L is the lag operator. There are many identifying conditions for these models, see for example Hamilton (1994) or Brockwell and Davis (2013).

Or more generally, we can use a nonlinear function of lag operators,

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

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

2.4.2 Fused Time-Series/Panel Models

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

$$\begin{array}{l} Y_{1t}^N = C_t^N + \varepsilon_t \\ Y_{1t}^I = C_t^N + \alpha_t + \varepsilon_t \end{array} \quad \Big| \quad \begin{array}{l} \varepsilon_t = \rho(\varepsilon_{t-1}) + u_t, \{u_t\} \text{ i.i.d. } E(u_t) = 0, \\ \{u_t\} \text{ is independent of } \{C_t^N\}, \end{array} \quad \Big| \quad t = 1, \ldots, T, \tag{11}$$

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

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

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

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

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 we then we can identify the time-series structure off the residuals $Y_{1t}^N - C_t^N$, where the missing observations Y_{1t}^N for $t > T_0$ are obtained as $Y_{1t}^N = Y_{1t} - \alpha_t$. Estimation can proceed analogously. This can improve upon the quality of our inferential procedure.

3 Theory

In this section, we provide theoretical justification for the validity of our conformal inference method. We derive theoretical results that are non-asymptotic in nature and hence hold in *finite samples*. When strong assumptions are imposed, the proposed approach is exact in a model-free manner. Under very weak assumptions, finite-sample bounds are provided for the size properties of our procedure; these bounds imply that our approach is asymptotically exact.

3.1 Exact Validity under Strong Assumptions

The following result shows that our conformal inference approach achieves finite sample size control if the estimated residuals $\{\hat{u}_t\}$ are exchangeable. The result is model-free in the sense that we do not need to use a correct or consistent estimator $\hat{P}_t^N$ for P_t^N.

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

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

Theorem 1 is the first main result of this paper. It states that if the residuals are exchangeable, under the null, the proposed conformal inference method achieves finite sample size control. Exchangeability of the residuals is implied, for example, if the data $\{Z_t\}_{t=1}^T$ are i.i.d. under the null, as shown in Lemma 1, but holds more generally. For example, in the difference-in-difference model the outcomes data can have an arbitrary common trend eliminated by differencing, making it possible for $\hat{u}_t = \hat{P}_t^N - P_t^N$ to be i.i.d. (or exchangeable more generally) with non i.i.d. data.

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

Of course, the exchangeability assumption is strong and may not be plausible in many applications. However, it allows us to discipline the choice of our inference procedure. Any permutation procedure which approximately works under dependence should have desirable properties under exchangeability. Our procedure enjoys exact finite sample validity and is fully robust to misspecification of the method for estimating P_t^N.

3.2 Approximate Validity under Weak Assumptions

In this section, we show that the proposed inference procedure has an approximate justification when the residuals are not exchangeable.

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

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

Assumption 2 is the main condition underlying our results. We can view it as much weaker than the previous assumption, since the data can be very general. Assumption 2.1 of i.i.d. shocks is our first sufficient condition. Under this condition, we will be able to use i.i.d. permutations, giving us a precise estimate of the p-value. The i.i.d. assumption can be replaced by Assumption 2.2, which is a widely accepted, weak condition, holding for many commonly encountered stochastic processes. It can be easily replaced by an even weaker ergodicity condition, as can be inspected in the proofs. Under this assumption, we will have to rely on the moving block permutations.

Remark 2. The assumption above can be generalized further, by requiring that the stochastic process $\{u_t\}_{t=1}^T$ satisfies one of the following conditions conditional on a random element V:

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

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

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

(1) the mean squared estimation error is small, $\|\hat{P}^N - P^N\|_2^2/T \leq \delta_T^2$;

(2) for $T_0 + 1 \leq t \leq T$, the pointwise errors are small, $|\hat{P}_t^N - P_t^N| \leq \delta_T$;

Assumption 3 imposes weak and easy-to-verify conditions on the performance of the estimators $\hat{P}_t^N$ of the counterfactual mean proxies P_t^N. These conditions are readily implied by the existing results for many estimators discussed in Section 2. In Section 5, we provide explicit conditions and references to explicit conditions, which imply these conditions.

Theorem 2 (**Approximate Validity of the Conformal Inference for Policy Effects**)**.** *We assume that T_* is fixed, and $T \to \infty$. Suppose that the Counterfactual Model stated in Assumption 1 holds, and that Assumption 3 holds. Impose Assumption 2.1 if i.i.d. permutations Π are used. Impose Assumption 2.2, if moving block permutations are used. Assume the statistic $S(u)$ has a density function bounded by D under the null. Then under the null hypothesis H_0, the p-value is approximately unbiased in size:*

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

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

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

4 Extensions

In this section, we discuss two extensions of our main results.

4.1 Testing Hypotheses about Average Effects

In addition to testing sharp null hypotheses, researchers are often also interested in testing hypotheses about average effects (e.g., Gobillon and Magnac, 2016; Carvalho et al., 2017; Li, 2017):

$$H_0 : \bar{\alpha} = \bar{\alpha}^o, \tag{12}$$

where

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

For any random variable V_t, let $\bar{V}_r = T_*^{-1} \sum_{t=r}^{r+T_*-1} V_t$. To simplify the exposition, we assume that T/T_* is an integer. Our inference procedure can be modified to test hypothesis (12), provided that there exists a model for the average counterfactual proxies $\bar{P}_r^N$:

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

where $\{\bar{u}_r\}$ is a stationary sequence. Our key assumption is that $\bar{P}_r^N$ can be identified and estimated based on the aggregated data:

$$\{\bar{Y}_{1r}, \ldots, \bar{Y}_{J+1r}, \bar{X}_{1r}, \ldots, \bar{X}_{J+1r}\}_{r=1}^{T_0+1}.$$

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

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

Note that testing hypothesis (12) is equivalent to testing the simple hypothesis (2) based on the aggregated data $\bar{Z}$. We compute $\hat{\bar{P}}_r^N$ based on $\bar{Z}$ and compute the residuals

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

The test statistic is

$$S\left(\hat{\bar{u}}\right)$$

and p-values can be computed based on permutations of $(\hat{\bar{u}}_1, \hat{\bar{u}}_{T_*+1}, \ldots, \hat{\bar{u}}_{T_0+1})'$ as described in Section 2.2. The finite sample and asymptotic properties of this test follow immediately from the results in Section 3.

4.2 Specification Tests

Here, we propose an easy-to-implement specification test for the key condition underlying our procedure: Assumption 1 (CFM). The key testable implication of Assumption 1 is stationarity of $\{u_t\}$. Consider the following null hypothesis:

$$H_0 : u_1 \overset{d}{=} u_{T_0},\ \{u_t\}_{t=0}^{T_0-1} \text{ is a stationary sequence.} \tag{13}$$

The conformal inference procedure developed in Section 2 naturally allows for testing hypothesis (13). Based on an appropriate method, we compute $\hat{P}_t^N$ using the pre-treatment data $\{Z_t\}_{t=1}^{T_0}$ and obtain the residuals

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

A natural test statistic is

$$S(\hat{u}) = |\hat{u}_{T_0}|,$$

which is constructed such that high-values indicate rejection of the null hypothesis. p-values can be computed based on permutations of $(\hat{u}_1, \ldots, \hat{u}_{T_0})'$ as described in Section 2.2. The finite sample and asymptotic properties of this specification test follow directly from the results in Section 3.

5 Verifying Small Estimation Error for Specific Models of Counterfactual Proxies

In this section, we revisit the models of counterfactual proxies introduced in Section 2. Primitive conditions are provided to guarantee that the estimation of counterfactual proxies is accurate enough for the validity of the proposed procedure. In particular, these conditions can be used to verify Assumption 3. In contrast to Section 2, we impose the null and estimate P_t^N using all T periods.

5.1 Difference-in-Differences

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

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

We consider the following estimator for the counterfactual:

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

where

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

Since $\hat{P}_t^N - P_t^N = \hat{\mu} - \mu$, Assumption 3 holds for the simple difference-in-differences model provided that $T^{-1}\sum_{t=1}^{T} u_t = o_P(1)$, which is true under very weak conditions.

5.2 Synthetic Control and Constrained Least-Squares Estimators

Several models in Section 2 (including synthetic control and constrained least-square methods) imply a structure in which the counterfactual proxy is a linear function of observed outcomes of untreated units.

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

$$Y = Xw + u, \tag{14}$$

where $u = (u_1, u_2, \ldots, u_T)'$ and $T = T_0 + T_*$. Identification is achieved by requiring that X and u be uncorrelated (cf. condition (SC)).

Under the framework in (14), different models correspond to different specifications for the weight vector w. For the synthetic control 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 synthetic control estimator

$$\hat{w} \quad = \quad \arg\min_w \ \|Y - Xw\|_2 : \ \text{s.t. } w \in \mathcal{W} = \{v \geq 0, \|v\|_1 = 1\}. \tag{15}$$

Moreover, we study the constrained Lasso estimator

$$\hat{w} = \arg\min_w \|Y - Xw\|_2 : \ \text{s.t. } w \in \mathcal{W} = \{v : \|v\|_1 \leq K\} \tag{16}$$

where $K > 0$ is a tuning parameter. In light of the estimator (15), 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-square estimator is consistent and satisfies Assumption 3.[8]

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

$$\hat{w} = \arg\min_w \ \|Y - Xw\|_2 : \ \text{s.t. } w \in \mathcal{W},$$

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

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

Lemma 2 provides some features that are important for counterfactual inference in our setup. First, we allow J to be large relative to T. To be precise, we only require $\log J = o(T^c)$, where

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

$c > 0$ is a constant depending only on the β-mixing coefficients; see Appendix for details. This is particularly relevant for problems in which the the number of (potential) control units and the number of time periods have similar order of magnitude; see for instance the applications in Abadie et al. (2010), Abadie et al. (2015), and Peri and Yasenov (2015). It is also important to note that the result in Lemma 2 does not require any sparsity assumptions on w, allowing for dense vectors. Moreover, compared to typical high-dimensional estimators (e.g., Lasso or Dantzig selector), our estimator does not require tuning parameters that can be difficult to choose in practice.

5.3 Factor models

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

5.3.1 Pure Factor/Matrix Completion Models

Recall from Section 2.3.4 the standard large factor model

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

where $F = (F_1, \ldots, F_T)' \in \mathbb{R}^{T \times k}$ 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 (Condition FE).

We use the standard principal component analysis (PCA) for estimating P_t^N. 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 result provides a theoretical guarantee on the estimation error.

Lemma 3 (Factor/Matrix Completion Model)**.** *Assume standard regularity conditions given in Bai (2003) including the identification condition FE, listed at the beginning of the proof of this lemma. Consider the factor model and the principal component estimator. Then for any $1 \leq t \leq T$, as $N \to \infty$ and $T \to \infty$*

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

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

5.3.2 Factor plus Regression Model: Interactive Fixed Effects Model

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

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

where $X_{jt} \in \mathbb{R}^{k_x}$ is observed covariates and $F = (F_1, \ldots, F_T)' \in \mathbb{R}^{T \times k}$ and $\Lambda = (\lambda_1, \ldots, \lambda_N)' \in \mathbb{R}^{N \times k}$ represent the k-dimensional unobserved factors and their loadings, respectively. The counterfactual proxy for Y_{1t}^N is $P_t^N = \lambda_1' F_t + X_{1t}'\beta$. In this model, we identify the counterfactual proxy through the condition that the idiosyncratic terms are independent of the factor structure and the observed covariates.

The two most popular estimation strategies are common correlated effects (CCE) estimators by Pesaran (2006) and PCA (the least squares) estimators by Bai (2009). In this paper, we follow the least squares approach but analogous results for CCE estimators can be established. 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 to our general methodology proposed in Section 2.

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

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

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

5.4 Time Series and Fused Models

As pointed out in Section 2.4, time series models, such as AR models, can be used to model the counterfactual proxy 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 general conformal inference approach.

5.4.1 AR Models

Recall from Section 2.4 the autoregressive models for the outcome with K lags:[9]

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

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

The estimation for P_t^N follows the ordinary least-square principle. Let

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

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

where $\hat{\rho}$ is the least squares estimators

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

The natural estimator for P_t^N is simply $\hat{P}_t^N = y_t' \hat{\rho}$. We implement the permutation based on $\hat{u}_t = Y_{1t}^N - \hat{P}_t^N$.

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

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

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

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

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

5.4.2 Fused Panel/Time Series Models with AR Errors

Here, we provide generic conditions for fused panel/time series models described in Section 2.4. In particular, AR models can be used to filter the estimated residuals and obtain near i.i.d errors. In Equation (11) of Section 2.4, we introduce an autoregressive structure in the error terms:

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

where C_t^N can be specified as a panel data model discussed before. Due to the autoregressive structure in ε_t, the counterfactual proxy is $P_t^N = C_t^N + \rho(\varepsilon_{t-1})$.

The estimation for P_t^N is done via a two-stage procedure. In the first stage, we estimate C_t^N using the techniques we considered before and obtain say $\hat{C}_t^N$. In the second stage, we estimate $\rho(\varepsilon_{t-1})$ by fitting the estimated residuals $\{\hat{\varepsilon}_t\}_{t=1}^T$ to an autoregressive model, where $\hat{\varepsilon}_t = Y_{1t}^N - \hat{C}_t^N$. For simplicity, we consider a linear model in the second stage estimation but analogous results can be obtained for more general models. To be specific, assume that

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

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

Given $\{\hat{\varepsilon}_t\}_{t=1}^T$ from the first-stage estimation, we define $\hat{x}_t = (\hat{\varepsilon}_{t-1}, \hat{\varepsilon}_{t-2}, ..., \hat{\varepsilon}_{t-K})' \in \mathbb{R}^K$ 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 i.i.d sequence with $E(u_t) = 0$ and $E(u_1^4)$ uniformly bounded and the roots of $1 - \sum_{j=1}^K \rho_j L^j = 0$ are uniformly bounded away from the unit circle. We assume that (1) $\sum_{t=1}^T (\hat{C}_t^N - C_t^N)^2 = o_P(T)$, (2) $\hat{C}_t^N - C_t^N = o_P(1)$ for $T_0 - K + 1 \leq t \leq T$. Then $\sum_{t=K+1}^T \left(\hat{P}_t^N - P_t^N\right)^2 = o_P(T)$ and $\hat{P}_t^N - P_t^N = o_P(1)$ for $T_0 + 1 \leq t \leq T$.*

Notice that the conditions in Lemma 7 for the autoregressive part are the same as in Lemma 5. The requirement on the consistency of $\hat{C}_t^N$ can be verified using existing results, e.g., those in Sections 5.1–5.3.

6 Simulations

This section presents simulation evidence on the finite sample properties of our inference procedure. Our simulation design is similar to Hahn and Shi (2016). The control outcomes are generated using a factor structure:

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

where $\mu_j = j/J$, $\lambda_j = j/J$, $\theta_t \overset{iid}{\sim} N(0,1)$, $F_t \overset{iid}{\sim} N(0,1)$, and $\epsilon_{jt} = \rho_\epsilon \epsilon_{jt-1} + \xi_{jt}$, $\xi_{jt} \overset{iid}{\sim} N(0, 1-\rho_\epsilon^2)$. We consider two different data generating processes (DGPs) for the treated unit. DGP1 specifies the outcome of the treated unit as a weighted average of the control units:

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

where $u_t = \rho_u u_{t-1} + v_t$, $v_t \overset{iid}{\sim} N(0, 1-\rho_u^2)$. The weights are either sparse

$$w = (0.5, 0.3, 0.15, 0.05, 0, \ldots, 0)' \quad \text{(DGP1a)}$$

or dense

$$w = (1/J, \ldots, 1/J)' \quad \text{(DGP1b)}.$$

Under DGP2, the treated outcome is generated by a factor structure:

$$Y_{1t}^N = \mu_1 + \theta_t + \lambda_1 F_t + u_t,$$

where $u_t = \rho_u u_{t-1} + v_t$, $v_t \overset{iid}{\sim} N(0, 1-\rho_u^2)$. The unit specific fixed effects and the factor loading are set such that there is common support between the treated and the control units in which case $\mu_1 = \lambda_1 = 0.5$ (DGP2a) or such that there is no common support in which case $\mu_1 = \lambda_1 = -0.5$ (DGP2b). For all DGPs we vary ρ_u, ρ_ϵ, T_0, and J.

We analyze the simple hypothesis testing problem (2) with $T_* = 1$ based on moving block permutations and consider three different approaches for estimating the counterfactual mean proxies P_t^N: (i) synthetic control (Section 2.3.2), (ii) a factor model without covariates (Section 2.3.4), and (3) constrained Lasso (Section 2.3.2). Synthetic control and constrained Lasso are correctly specified for DGP1 whereas the factor model is correctly specified with DGP2.

The simulation results reported in Tables 1 and 2 confirm our theoretical results. If the data are i.i.d. (implying exchangeability of the residuals as shown in Lemma 1), our procedure achieves exact

size control, irrespectively of whether the method used to estimate the counterfactual is correctly specified or not. With dependent data, the proposed procedure exhibits close-to-correct size, even when the model for P_t^N is misspecified.

To investigate the power of our inference procedures, we consider a fixed alternative of $\alpha_T = 2$. The results in Tables 3 and 4 demonstrate that our procedure enjoys good power properties. Power is tends to be high for correctly specified models but can be substantially lower for misspecified models; see example the synthetic control model under DGP2b.

7 Empirical Illustration

In this section, we apply our inference procedure to analyze the effect of election day registration (EDR) laws on voter turnout in the United States as in Xu (2017). Voting in the United States is typically a two-step procedure since eligible voters must register prior to casting their ballots. Registration usually requires a separate trip, which imposes additional costs on voters and therefore potentially leads to low turnout rates. EDR is a reform that allows eligible voters to register on the election day when arriving at the polling stations. In the mid 1970s, Maine Minnesota, and Wisconsin adopted this reform. Idaho, New Hampshire, and Wyoming introduced EDR in the 1990s and Montana, Iowa, and Connecticut enacted EDR before the 2012 presidential election.

We use state-level voter turnout data for presidential elections from 1920 to 2012, previously analyzed in Xu (2017) to which we refer for more information about the dataset and descriptive statistics. Turnout rates are computed by dividing total ballots counted by the state's voting-age population. Alaska and Hawaii are excluded because they were no states until 1959 and North Dakota is excluded as no voter registration is needed there. We analyze each of the nine treated state separately.[10] The $J = 38$ states which did not enact EDR laws between 1920 and 2012 serve as control units. Since the EDR laws were enacted in three waves, the number of pre- and post-treatment periods (T_0, T_*) differs across states. For the first wave (Maine, Minnesota, Wisconsin), $(T_0, T_*) = (14, 10)$; for the second wave (Idaho, New Hampshire, and Wyoming), $(T_0, T_*) = (19, 5)$; and for the third wave $(T_0, T_*) = (22, 2)$ for Montana and Iowa and $(T_0, T_*) = (23, 1)$ for Connecticut. Figure 2 displays the raw turnout data for treated and control states.[11]

We consider three different methods for estimating the counterfactual mean proxy P_t^N: (i) canonical synthetic control (Section 2.3.2), (ii) a pure factor model without covariates and two factors (Section 2.3.4), and (iii) constrained Lasso (Section 2.3.2).

We first test the no-effects null hypothesis

$$H_0 : (\alpha_{T_0+1}, \ldots, \alpha_T)' = (0, \ldots, 0)'. \tag{17}$$

Note that the underlying hypotheses differ by state because the number of post-treatment periods T_* differs across states. Table 5 reports p-values based on moving block and i.i.d. permutations.[12]

The results differ substantially across the different methods for estimating counterfactual proxies. For synthetic control, we can reject the null hypothesis (17) for the majority of the states, while we only reject the null for very few states for the factor model and constrained Lasso. While i.i.d. permutations often yield slightly lower p-values than moving block permutations, the substantive overall conclusions are not affected by the choice of the set of permutations.

[10] Xu (2017) provides a state-by-state analysis in the online supplemental material.

[11] This figure is an adapted version of Figure A5 in the supplementary material to Xu (2017).

[12] To keep estimation tractable, we use a random subset of 5000 permutations.

Figures 3–5 display pointwise confidence intervals for the policy effect α_t based on test inversion and moving block permutations. There is substantial heterogeneity in the effect of the EDR laws on turnout rates across states. For example, for New Hampshire, we find significantly positive effects for many periods and for all three methods. In contrast, for Connecticut, there are negative effects for synthetic control and constrained Lasso and no significant effect for the factor model, while for other states such as Montana, EDR laws do not significantly impact turnout in any period. These findings are broadly consistent with the results in Xu (2017). As for the overall no-effects null hypothesis, the choice of the model for the counterfactual matters for the confidence intervals.

8 Conclusion

This paper introduces new inference procedures for counterfactual and synthetic control methods for evaluating policy effects. Our procedures work in conjunction with a great variety of powerful methods for estimating the counterfactual mean outcome in the absence of a policy intervention. The proposed approach has a double justification, in that the inference result is exact under strong assumptions on data, and is approximately exact under very weak assumptions on the data. Weak and easy-to-verify conditions are provided for methods that can be used to estimate the counterfactual, allowing for temporally and cross-sectionally dependent data. The new approach demonstrates an excellent performance in simulation experiments, and is taken to a data application, where we re-evaluate the causal effect of election day registration (EDR) laws on voter turnout in the United States.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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., Xu, L., et al. (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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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. (2017). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, (just-accepted).

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

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

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.

Peri, G. and Yasenov, V. (2015). The labor market effects of a refugee wave: Applying the synthetic control method to the mariel boatlift. Working Paper 21801, National Bureau of Economic Research.

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

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

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

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

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.

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

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

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

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

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. (2014). *Asymptotic theory for econometricians*. Academic press.

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

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

Notations

We introduce some notations that will be used in the rest of the paper. Let $\mathbb{Z}$ denote the set of integers. For any $a \in \mathbb{R}$, we define $\lfloor a \rfloor = \max\{z \in \mathbb{R} : z \leq a\}$ and $\lceil a \rceil = \min\{z \in \mathbb{Z} : z \geq a\}$. For $a, b \in \mathbb{R}$, $a \vee b = \max\{a, b\}$. For a set A, $|A|$ denotes the cardinality of A. For two positive sequences a_n, b_n, we use $a_n \ll b_n$ to denote $a_n = o(b_n)$; $a_n \lesssim b_n$ means that there exists a universal constant $C > 0$ with $a_n \leq C b_n$. Moreover, $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \lesssim a_n$. 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.

A General Results on Exact and Approximate Randomization Inference

We present several results that can be of independent interest in many conformal/randomization inference problems.

Setting. Let $\hat{u} = \{\hat{u}_t\}$ and $u = \{u_t\}$ arbitrary stochastic process indexed by $t \in \{1, ..., T\}$ taking values in a sample space $\mathcal{U}_T$. We regard $\hat{u}$ as an estimator for u in the asymptotic results below.

Let $\hat{u}^\pi = \{\hat{u}^\pi(t)\}$ and $u^\pi = \{u_t^\pi\}$ with $\pi \in \Pi$ be an indexed collection of arbitrary stochastic processes indexed by $t \in \{1, ..., T\}$ taking values in $\mathcal{U}_T$. We regard these processes as randomized versions of $\hat{u}$ and u. We assume that the index Π includes an identity element $\mathbb{I}$ so that $\hat{u} = \hat{u}^\mathbb{I}$ and $u = u^\mathbb{I}$.

There are two main examples of considerable interest to us.

1. **Permutation of Residuals:** Let π designate permutation of residuals, namely

$$\hat{u}^\pi = \{\hat{u}_{\pi(t)}\} \quad \text{and} \quad u^\pi = \{u_{\pi(t)}\}$$

   where $\pi \in \Pi$, a collection of one-to-one permutation maps on $\{1, ..., T\}$, including the identity map.

2. **Residuals Resulting from Permutation of Data:** Let π designate permutations on an underlying data frame, namely

$$\hat{u}^\pi = \{g(Z_{\pi(t)}, \hat{\beta}(\{Z_{\pi(t)}\}_{t=1}^T))\} \quad \text{and} \quad u^\pi = \{g(Z_{\pi(t)}, \beta_0)\},$$

   where $\pi \in \Pi$, a collection of one-to-one permutation maps on $\{1, ..., T\}$, including the identity map, where $\{Z_t\}_{t=1}^T$ is the data frame taking values in the sample space $\mathcal{Z}_T$ and $\hat{\beta} : \mathcal{Z}_T \to \mathcal{B}_T$ is a measurable estimator map from the sample space to the parameter space, and $g : \mathcal{Z}_T \times \mathcal{B}_T \to \mathcal{U}_T$ is a measurable map.

Exact Validity. Let $\{S^{(j)}(\hat{u})\}_{j=1}^n$ denoted the non-decreasing rearrangement of $\{S(\hat{u}^\pi) : \pi \in \Pi\}$, where $n = |\Pi|$. Call these randomization quantiles. Define the randomization p-value:

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

Observe that

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

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

Proposition 1 (General Exact Validity). *Suppose that* $\{\hat{u}_t^\pi\}$ *has an exchangeable distribution under* $\pi \in \Pi$. *Consider any fixed* Π *such that the randomization* α*-quantiles are invariant surely, namely*

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

Or, more generally, suppose that surely

$$S^{(k(\alpha))}(\hat{u}^\pi) \geq S^{(k)}(\hat{u}), \text{ for all } \pi \in \Pi. \tag{18}$$

Then

$$P(\hat{p} \leq \alpha) = P(S(\hat{u}) > S^{(k)}(\hat{u})) \leq \alpha.$$

Approximate Validity. For approximate results, assume that the number of randomizations becomes large, $n = |\Pi| \to \infty$ (in examples above, this is caused by $T \to \infty$). Let $\{\delta_{1n}, \delta_{2n}, \gamma_{1n}, \gamma_{2n}\}$ be sequences of numbers converging to zero, and assume the following conditions.

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

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

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

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

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

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

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

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

Consider the approximate randomization p-value $\hat{p} = 1 - \hat{F}(S(\hat{u}))$.

Proposition 2 (**Approximate General Validity of the Randomization/Conformal Inference**). *Under the approximate ergodicity condition (E) and the small error condition (A), the approximate conformal p-value obeys for any* $\alpha \in (0, 1)$

$$|P\left(\hat{p} \leq \alpha\right) - \alpha| \leq 3\delta_{1n} + 2(\delta_{2n} + D\sqrt{\delta_{2n}}) + \gamma_{1n} + \gamma_{2n}.$$

The theorem can be seen as a generalization of Hoeffding (1952) result in that it is non-asymptotic, not requiring the convergence in distributions of relevant statistics, which is the case in our setting. It is also stated in terms of "estimated residuals" and their closeness to the true residuals, making it easy to apply in a variety of problems. For example, they can be used to justify the permutation inference procedure proposed by Abadie et al. (2010, 2015) (with t exchanged by j).

A.1 Proof of Proposition 1

We have by (18)

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

Since $\mathbf{1}(S(\hat{u}) > S^{(k)}(\hat{u}))$ is equal in law to $\mathbf{1}(S(\hat{u}^\pi) > S^{(k)}(\hat{u}^\pi))$ for any $\pi \in \Pi$ by the exchangeability hypothesis, then

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

A.2 Proof of Proposition 2

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,

$$P(\mathcal{M}) \geq 1 - \gamma_{1n} - \gamma_{2n}. \tag{19}$$

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

$$\begin{aligned}
\left|\hat{F}(S(\hat{u})) - F(S(u))\right| &\leq \left|\hat{F}(S(\hat{u})) - F(S(\hat{u}))\right| + |F(S(\hat{u})) - F(S(u))| \\
&\overset{\text{(i)}}{\leq} \sup_{x\in\mathbb{R}}\left|\hat{F}(x) - F(x)\right| + D\,|S(\hat{u}) - S(u)| \\
&\leq \sup_{x\in\mathbb{R}}\left|\hat{F}(x) - \tilde{F}(x)\right| + \sup_{x\in\mathbb{R}}\left|\tilde{F}(x) - F(x)\right| + D\,|S(\hat{u}) - S(u)| \\
&\leq \sup_{x\in\mathbb{R}}\left|\hat{F}(x) - \tilde{F}(x)\right| + \delta_{1n} + D\,|S(\hat{u}) - S(u)| \\
&\leq \sup_{x\in\mathbb{R}}\left|\hat{F}(x) - \tilde{F}(x)\right| + \delta_{1n} + D\delta_{2n},
\end{aligned} \tag{20}$$

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

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

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

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

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

$$\sup_{x\in\mathbb{R}}\left|\hat{F}(x) - \tilde{F}(x)\right| \leq 2\delta_{1n} + 2\delta_{2n} + D\sqrt{\delta_{2n}}. \tag{21}$$

We combine (20) and (21) and obtain that on the event $\mathcal{M}$,

$$\left|\hat{F}(S(\hat{u})) - F(S(u))\right| \leq 3\delta_{1n} + 2\delta_{2n} + 2D\sqrt{\delta_{2n}}. \tag{22}$$

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

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

where (i) follows by the elementary inequality $|\mathbf{1}\{1 - \hat{F}(S(\hat{u})) \leq \alpha\} - \mathbf{1}\{1 - F(S(u)) \leq \alpha\}| \leq \mathbf{1}\{|F(S(u)) - 1 + \alpha| \leq |\hat{F}(S(\hat{u})) - F(S(u))|\}$, (ii) follows by (22), (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 (19). The proof is complete.

B Proofs of Results Stated in the Main Text

B.1 Proof of Lemma 1

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

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

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

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

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

B.2 Proof of Theorem 1

The result of the theorem follows because the Π considered all obey $\Pi\pi = \Pi$ for all $\pi \in \Pi$. The result then follows from a general theorem in permutation inference given in the first page of Romano (1990)'s article, or from Proposition 1.

B.3 Proof of Theorem 2

The result is a consequence of the following four lemmas, that verify the approximate ergodicity conditions (E) and conditions on the estimation error (A) of Proposition 2. 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 8 (Mixing Implies Approximate Ergodicity)**.** *Let Π be the moving block permutations. Suppose that $\{u_t\}_{t=1}^T$ is stationary and strong mixing. Assume the following conditions: (1) $\sum_{k=1}^{\infty} \alpha_{mixing}(k)$ is bounded by a constant M, and (2) $T_0 \geq T_* + 2$. Then there exists a constant $M' > 0$ depending only on M such that for any $\delta_{1n} > 0$,*

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

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

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

Lemma 9 (Approximate Ergodicity under I.I.D. Permutations)**.** *Let Π be the set of all permutations. Suppose that $\{u_t\}_{t=1}^T$ is i.i.d. Assume that $S(u)$ only depends on the last T_* entries of u. If $T_0 \geq T_* + 2$, then*

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

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

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

Lemma 10 (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$D_{T_*} :=$ Blockdiag$(0_{T_*}, I_{T_*})$. Then Condition (A) (1)-(2) is satisfied if there exist sequences $\gamma_n, \delta_{2n} = o(1)$ such that with probability at least $1 - \gamma_n$,*

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

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

Lemma 11 (Bounds on Estimation Errors under I.I.D. Permutations)**.** *Consider the set of all permutations Π. Let T_* be fixed. Suppose that for some constant $Q > 0$, $|S(u) - S(v)| \leq Q\|D_{T_*}(u - v)\|_2$ for any $u, v \in \mathbb{R}^T$ and $D_{T_*} :=$ Blockdiag$(0, I_{T_*})$. Then Condition (A) (1)-(2) is satisfied if there exist sequences $\gamma_n, \delta_{2n} = o(1)$ such that with probability at least $1 - \gamma_n$,*

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

Now we conclude the proof of Theorem 2.

For the moving block permutations, let $\delta_{1n} = (T_*/T_0)^{1/4}$. Then we apply Proposition 2 together with Lemmas 8 and 10, obtaining

$$\begin{aligned}
|P(\hat{p} \leq \alpha) - \alpha| &\leq 3\delta_{1n} + 2(\delta_{2n} + D\sqrt{\delta_{2n}}) + \gamma_{1n} + \gamma_{2n} \\
&\leq 3\delta_{1n} + 2(\delta_{2n} + D\sqrt{\delta_{2n}}) + \left(M'\sqrt{\frac{T_*}{T_0}}\log T_0 + \frac{T_*+1}{T_0+T_*}\right)/\delta_{1n} + \gamma_{2n}
\end{aligned}$$

$$\leq 3(T_*/T_0)^{1/4} + 2(\delta_{2n} + D\sqrt{\delta_{2n}}) + \left(M'\sqrt{\frac{T_*}{T_0}}\log T_0 + \frac{T_*+1}{T_0+T_*}\right)(T_*/T_0)^{-1/4} + \gamma_{2n}.$$

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 i.i.d permutations, we also use $\delta_{1n} = (T_*/T_0)^{1/4}$. Then we apply Proposition 2 together with Lemmas 9 and 11, obtaining

$$\begin{aligned}
|P(\hat{p} \leq \alpha) - \alpha| &\leq 3\delta_{1n} + 2(\delta_{2n} + D\sqrt{\delta_{2n}}) + \gamma_{1n} + \gamma_{2n} \\
&\leq 3\delta_{1n} + 2(\delta_{2n} + D\sqrt{\delta_{2n}}) + \sqrt{2\pi/\lfloor T/T_*\rfloor}/\delta_{1n} + \gamma_{2n} \\
&\leq 3(T_*/T_0)^{1/4} + 2(\delta_{2n} + D\sqrt{\delta_{2n}}) + \sqrt{2\pi/\lfloor T/T_*\rfloor}(T_*/T_0)^{-1/4} + \gamma_{2n} \\
&\lesssim (T_*/T_0)^{1/4} + \sqrt{\delta_{2n}} + \gamma_{2n}.
\end{aligned}$$

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

B.3.1 Proof of Lemma 8

We define

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

It is straight-forward to verify that

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

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

$$E\left(\sup_{x\in\mathbb{R}} \left|\check{F}(x) - F(x)\right|^2\right) \leq \frac{1}{T_0}\left(1 + 4\sum_{k=0}^{T_0-1} \tilde{\alpha}_{\text{mixing}}(t)\right)\left(3 + \frac{\log T_0}{2\log 2}\right)^2. \tag{23}$$

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

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

Since $\sum_{k=1}^{\infty} \alpha_{\text{mixing}}(k)$ is bounded by M, it follows by (23) that

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

By Liapunov's inequality,

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

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

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

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

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

The desired result follows by Markov's inequality.

B.3.2 Proof of Lemma 9

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

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

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

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

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

Therefore,

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

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

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

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

Hence by Jensen's inequality

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

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

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

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

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

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

The desired result follows by Markov's inequality.

B.3.3 Proof of Lemma 10

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

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

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.

B.3.4 Proof of Lemma 11

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

$$\begin{aligned}\sum_{\pi\in\Pi}\left(\hat{u}_{\pi(t)} - u_{\pi(t)}\right)^2 &= \sum_{s=1}^{T}\sum_{\pi\in A_{t,s}}\left(\hat{u}_{\pi(t)} - u_{\pi(t)}\right)^2 \\ &= \sum_{s=1}^{T}\sum_{\pi\in A_{t,s}}(\hat{u}_s - u_s)^2 = \sum_{s=1}^{T}|A_{t,s}|\,(\hat{u}_s - u_s)^2 = (T-1)! \times \|\hat{u} - u\|_2^2. \qquad (24)\end{aligned}$$

Due to the Lipschitz property of $S(\cdot)$, we have for some Q that depends on q and T^*

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

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

where the penultimate equality follows by (24) 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.

B.4 Proof of Lemma 2

Let X_{jt} denote the (j,t) entry of the matrix $X\in\mathbb{R}^{T\times J}$. We assume the following conditions hold: (1) $E(u_tX_{jt})=0$ for $1\leq j\leq J$. (2) there exist constants $c_1,c_2>0$ such that $E|X_{jt}u_t|^2\geq c_1$ and $E|X_{jt}u_t|^3\leq c_2$ for any $1\leq j\leq J$ and $1\leq t\leq T$; (3) for each $1\leq j\leq J$, the sequence $\{X_{jt}u_t\}_{t=1}^T$ is β-mixing and the β-mixing coefficient satisfies that $\beta(t)\leq a_1\exp(-a_2t^\tau)$, where $a_1,a_2,\tau>0$ are constants. (4) there exists a constant $c_3>0$ such that $\max_{1\leq j\leq J}\sum_{t=1}^T X_{jt}^2u_t^2\leq c_3^2T$ with probability $1-o(1)$. (5) $\log J = o(T^{4\tau/(3\tau+4)})$ and $w\in\mathcal{W}$. (6) There exists a sequence $\ell_T>0$ such that $(X_t'\delta)^2\leq\ell_T\|X\delta\|_2^2/T$, for all $w+\delta\in\mathcal{W}$ with probability $1-o(1)$ for $T_0+1\leq t\leq T$ and (7) $\ell_TB_T\to0$ for $B_T = M[\log(T\vee J)]^{(2+2\tau)/(4\tau)}T^{-1/2}$.

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

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

$$\|X(\hat{w}-w)\|_2^2/T \leq B_T = M[\log(T\vee J)]^{(2+2\tau)/(4\tau)}T^{-1/2}$$

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

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

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

Lemma 12. *Suppose that (1) $E(u_tX_{jt})=0$ for $1\leq j\leq J$. (2) $\max_{1\leq j\leq J,1\leq t\leq T}E|X_{jt}u_t|^3\leq K_1$ for a constant $K_1>0$. (3) $\min_{1\leq j\leq J,1\leq t\leq T}E|X_{jt}u_t|^2\geq K_2$ for a constant $K_2>0$. (4) For each $1\leq j\leq J$, $\{X_{jt}u_t\}_{t=1}^T$ is β-mixing and the β-mixing coefficients satisfy $\beta(s)\leq D_1T\exp(-D_2s^\tau)$ for some constants $D_1,D_2,\tau>0$. Assume $\log J = o(T^{4\tau/(3\tau+4)})$. Then there exists a constant $\kappa>0$ depending only on K_1,K_2,D_1,D_2,τ such that with probability $1-o(1)$*

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

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

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

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

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 $0 \leq x \leq C_0 k^{2/3}$,

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

Therefore, for any $0 \leq x \leq C_0 k^{2/3}$,

$$\begin{aligned} P\left(\left|\frac{\sum_{t\in H_i} W_{j,t}}{\sqrt{\sum_{t\in H_i} W_{j,t}^2}}\right| > x\right) &\leq P\left(\left|\frac{\sum_{t\in H_i}\tilde{W}_{j,t}}{\sqrt{\sum_{t\in H_i}\tilde{W}_{j,t}^2}}\right| > x\right) + P\left(\bigcup_{t\in H_i}\{\tilde{W}_{j,t} \neq W_{j,t}\}\right) \\ &\leq C_1\left(1-\Phi(x)\right) + k\beta(m). \quad (25) \end{aligned}$$

The Cauchy-Schwarz inequality implies

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

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 $0 \leq x \leq C_0 k^{2/3}\sqrt{m}$,

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

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

where (i) follows by (25) and (ii) follows by the inequality $1-\Phi(a) \le a^{-1}\phi(a)$ (with ϕ being the pdf of $N(0,1)$) and $\beta(m) \le D_1 \exp(-D_2 m^\tau)$.

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

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

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

If T/k is not an integer, then we simply add one observation from $\{W_{j,t}\}_{t=km+1}^T$ to each of H_i for $1 \le i \le m$. The bound in (25) holds with C_1 large enough. The proof is complete. □

Now we are ready to prove Lemma 2.

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

$$\|X\Delta\|_2^2 \le 2u'X\Delta \overset{\text{(i)}}{\le} 2\|Xu\|_\infty \|\Delta\|_1 \overset{\text{(ii)}}{\le} 4K\|Xu\|_\infty, \tag{26}$$

where (i) follows by Holder's inequality and (ii) follows by $\|\Delta\|_1 \le 2K$ (since $\|\hat{w}\|_1 \le K$ and $\|w\|_1 \le K$). By Lemma 12, there exists a constant $\kappa > 0$ such that

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

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

$$P\left(\max_{1\le j\le J}\left|\sum_{t=1}^T X_{jt}u_t\right| > \kappa c_3[\log(T\vee J)]^{(1+\tau)/(2\tau)}\sqrt{T}\right) = o(1). \tag{27}$$

Part (1) follows by combining (26) and (27). Part (2) follows by part (1) and $\log J = o(T^{\tau/(\tau+1)})$. □

B.5 Proof of Lemma 3

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

Suppose that there exists a constant $D_0 > 0$ the following conditions hold: (1) $\max_{1\le t\le T} E\|F_t\|^4 \le D_0$, $\max_{1\le j\le N}\|\lambda_j\|^4 \le D_0$, $\max_{jt} E|u_{jt}|^8 \le D_0$ and $E(u_{jt}) = 0$. (2) $\max_s N^{-1}\sum_{t=1}^T \left|\sum_{i=1}^N E(u_{is}u_{it})\right| \le D_0$ and $\max_i \sum_{j=1}^N \max_{1\le t\le T}|E(u_{it}u_{jt})| \le D_0$. (3) $(NT)^{-1}\sum_{s=1}^T\sum_{t=1}^T\sum_{i=1}^N\sum_{j=1}^N |E(u_{it}u_{js})| \le D_0$ and $\max_{s,t} E\left|N^{-1/2}\sum_{i=1}^N [u_{is}u_{it} - E(u_{is}u_{it})]\right|^4 \le D_0$. (4) $N^{-1}\sum_{i=1}^N E\left\|T^{-1/2}\sum_{t=1}^T F_t u_{it}\right\|^2 \le D_0$.

(5) $\max_t E\left\|(NT)^{-1/2}\sum_{s=1}^{T}\sum_{i=1}^{N}F_s[u_{is}u_{it}-E(u_{is}u_{it})]\right\|^2 \leq D_0$. (6) $E\left\|(NT)^{-1/2}\sum_{t=1}^{T}\sum_{i=1}^{N}F_t\lambda_i'u_{it}\right\|^2 \leq D_0$.

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 $\Phi_i = \lim_{T\to\infty}T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}E(F_tF_s'u_{is}u_{it})$. (9) $N^{-1}\sum_{i=1}^{N}\lambda_i\lambda_i' \to \Sigma_\Lambda$ and $T^{-1}\sum_{t=1}^{T}F_tF_t' = \Sigma_F + o_P(1)$ for some $k\times k$ positive definite matrices Σ_Λ and Σ_F satisfying that $\Sigma_\Lambda\Sigma_F$ has distinct eigenvalues.

What follows below is the proof of the lemma. We recall some notations used by Bai (2003). 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):

$$\hat{\lambda}_1'\hat{F}_t - \lambda_1'F_t = \left(\hat{F}_t - H'F_t\right)'H^{-1}\lambda_1 + \hat{F}_t'(\hat{\lambda}_1 - H^{-1}\lambda_1). \tag{28}$$

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),

$$\sum_{t=1}^{T}\|\hat{F}_t - H'F_t\|^2 = O_P(T/\delta_{NT}^2), \tag{29}$$

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

$$\sum_{t=1}^{T}\|\hat{F}_t\|^2 = \sum_{t=1}^{T}\mathrm{trace}(\hat{F}_t\hat{F}_t') = \mathrm{trace}(\hat{F}'\hat{F}) = kT. \tag{30}$$

By Theorem 2 of Bai (2003),

$$\hat{\lambda}_1 = H^{-1}\lambda_1 + O_P(\max\{T^{-1/2},N^{-1}\}). \tag{31}$$

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, $\|\lambda_1\| = O(1)$. Hence,

$$\|H^{-1}\lambda_1\| = O_P(1). \tag{32}$$

Step 2: prove the desired result.

Therefore,

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

where (i) follows by (28) and the elementary inequality of $(a+b)^2 \leq 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$ and (ii) follows by (29), (30), (31) and (32). Since $n = |\Pi| = T$ for moving block permutation, we have

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

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

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

B.6 Proof of Lemma 4

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

Suppose that there exists a constant $D_1 > 0$ the following conditions hold: (1) $\max_{i,t} E\|X_{it}\|^4 \leq D_1, \max_t E\|F_t\|^4 \leq D_1, \max_i E\|\lambda_i\|^4 \leq D_1$ 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 $E(u_iu_i')$ 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) Define $X_i = (X_{i1}, ..., X_{iT})' \in \mathbb{R}^{T\times k_x}$ and $M_F = I_T - F(F'F)^{-1}F'$, we have

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

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

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

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

$$\hat{\lambda}_1 = T^{-1}\hat{F}'(Y_1^N - X_1\hat{\beta}) = T^{-1}\hat{F}'(u_1 + F\lambda_1 - X_1\Delta_\beta). \tag{33}$$

Notice that

$$\begin{aligned}
|\hat{u}_{1t} - u_{1t}|^2 &= \left|F_t'\lambda_1 - \hat{F}_t'\hat{\lambda}_1 - X_{1t}'\Delta_\beta\right|^2 \\
&\overset{(i)}{=} \left|F_t'\lambda_1 - T^{-1}(H'F_t + \Delta_{F,t})'\hat{F}'(u_1 + F\lambda_1 - X_1\Delta_\beta) - X_{1,t}'\Delta_\beta\right|^2 \\
&\leq \left[\left|F_t'\left(I_k - H\hat{F}'F/T\right)\lambda_1\right| + \left|T^{-1}\Delta_{F,t}'\hat{F}'F\lambda_1\right| + \left|T^{-1}\hat{F}_t'\hat{F}'(u_1 - X_1\Delta_\beta)\right| + \left|X_{1t}'\Delta_\beta\right|\right]^2 \\
&\lesssim \left[F_t'\left(I_k - H\hat{F}'F/T\right)\lambda_1\right]^2 + \left[T^{-1}\Delta_{F,t}'\hat{F}'F\lambda_1\right]^2 + \left[T^{-1}\hat{F}_t'\hat{F}'(u_1 - X_1\Delta_\beta)\right]^2 + \left[X_{1t}'\Delta_\beta\right]^2,
\end{aligned} \tag{34}$$

where (i) follows by (33) and $\hat{F}_t = H'F_t + \Delta_{F,t}$. Therefore,

$$\begin{aligned}
\sum_{t=1}^{T}(\hat{u}_{1t}-u_{1t})^2 &\lesssim \sum_{t=1}^{T}\left[F_t'\left(I_k - H\hat{F}'F/T\right)\lambda_1\right]^2 + \sum_{t=1}^{T}\left[T^{-1}\Delta_{F,t}'\hat{F}'F\lambda_1\right]^2 \\
&\quad + \sum_{t=1}^{T}\left[T^{-1}\hat{F}_t'\hat{F}'(u_1 - X_1\Delta_\beta)\right]^2 + \sum_{t=1}^{T}\left[X_{1t}'\Delta_\beta\right]^2 \\
&\overset{\text{(i)}}{=} \lambda_1'\left(I_k - H\hat{F}'F/T\right)'(F'F)\left(I_k - H\hat{F}'F/T\right)\lambda_1 \\
&\quad + T^{-2}\left(\hat{F}'F\lambda_1\right)'(\Delta_F'\Delta_F)\left(\hat{F}'F\lambda_1\right) + T^{-1}\left\|\hat{F}'(u_1 - X_1\Delta_\beta)\right\|^2 + \|X_1\Delta_\beta\|^2 \\
&\overset{\text{(ii)}}{=} O_P\left(T\|\Delta_\beta\|^2 + T\delta_{NT}^{-4}\right) + O_P\left(T\|\Delta_\beta\|^2 + T\delta_{NT}^{-2}\right) + O_P\left(1 + T\delta_{NT}^{-4} + T\|\Delta_\beta\|^2\right) + O_P(T\|\Delta_\beta\|^2) \\
&= O_P\left(1 + T\|\Delta_\beta\|^2 + T\delta_{NT}^{-2}\right),
\end{aligned}$$

where (i) follows by $\sum_{t=1}^{T}\hat{F}_t\hat{F}_t' = \hat{F}'\hat{F} = TI_k$ and (ii) follows by Lemma 13, 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 $\|\Delta_\beta\| = O_P(1/\sqrt{NT}) = O_P(T^{-1})$. 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 (34), we have

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

where (i) follows by $I_k - H\hat{F}'F/T = O_P(\|\Delta_\beta\| + \delta_{NT}^{-2})$, $\|\Delta_F\| = O_P(\sqrt{T}\|\Delta_\beta\| + \sqrt{T}\delta_{NT}^{-1})$ and $\|\hat{F}'(u_1 - X_1\Delta_\beta)\| = O_P(\sqrt{T} + T\delta_{NT}^{-2} + T\|\Delta_\beta\|)$ (due to Lemma 13), whereas (ii) follows by $\|\hat{F}_t\| = O_P(1)$ (Lemma 13), $\|X_{1t}\| = O_P(1)$, $\|F_t\| = O_P(1)$, $\lambda_1 = O(1)$, $\|\Delta_\beta\| = O_P(T^{-1})$ and $\|F\lambda_1\| = O_P(\sqrt{T})$.

Lemma 13. *Suppose that the assumption of Theorem 4 holds. Let δ_{NT}, H, Δ_F and u_1 be defined as in the proof of Theorem 4. Then (1) $I_k - H\hat{F}'F/T = O_P(\|\Delta_\beta\| + \delta_{NT}^{-2})$; (2) $\Delta_F'\Delta_F = O_P(T\|\Delta_\beta\|^2 + T\delta_{NT}^{-2})$; (3) $\left\|\hat{F}'(u_1 - X_1\Delta_\beta)\right\| = O_P\left(\sqrt{T} + T\delta_{NT}^{-2} + T\|\Delta_\beta\|\right)$; (4) $\|X_1\Delta_\beta\| = O_P(\sqrt{T}\|\Delta_\beta\|)$; (5) $\hat{F}\Delta_{F,t} = O_P(T\|\Delta_\beta\| + T\delta_{NT}^{-2})$; (6) $\|\hat{F}_t\| = O_P(1)$ for $1 \le t \le T$.*

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

$$H = O_P(1) \quad \text{and} \quad H^{-1} = O_P(1). \tag{35}$$

Notice that

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

$$\overset{\text{(iii)}}{=} O_P(\|\Delta_\beta\|) + O_P(\delta_{NT}^{-2}), \tag{36}$$

where (i) holds by $\hat{F} = FH + \Delta_F$, (ii) holds by $I_k - (HH')(F'F/T) = O_P(\|\Delta_\beta\|) + O_P(\delta_{NT}^{-2})$ (due to Lemma A.7(i) in Bai (2009)) and (iii) holds by (35) and $\Delta_F'F/T = O_P(\|\Delta_\beta\|) + O_P(\delta_{NT}^{-2})$ (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):

$$T^{-1}\Delta_F'\Delta_F = O_P(\|\Delta_\beta\|^2) + O_P(\delta_{NT}^{-2}). \tag{37}$$

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

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

It follows that

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

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

$$\|F'u_1\| = O_P(\sqrt{T}). \tag{38}$$

Notice that

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

where (i) follows by $\left(\hat{F} - FH\right)' u_1/T = O_P(T^{-1/2}\|\Delta_\beta\|) + O_P(\delta_{NT}^{-2})$ (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 (35) and (38). 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 $\|X_1\Delta_\beta\| \leq \|X_1\| \cdot \|\Delta_\beta\|$.

Proof of part (5). Notice that

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

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

Proof of part (6). Notice that

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

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 $\|\Delta_{F,t}\| = O_P(1)$. By $\|\hat{F}_t\| \leq \|H'F_t\| + \|\Delta_{F,t}\|$, $F_t = O_P(1)$ and $H = O_P(1)$, we can see that $\|\hat{F}_t\| = O_P(1)$. The proof is complete. □

B.7 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,

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

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

$$T^{-1} \sum_{t=1}^{T} u_t^2 = E(u_t^2) + o_P(1),$$

which means $\sum_{t=1}^{T} u_t^2 = O_P(T)$. Since $\hat{\rho} - \rho = o_P(1)$, the above display implies that

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

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

B.8 Proof of Lemma 7

In this proof, we use $\|\cdot\|$ to denote the Euclidean norm of vectors or the spectral norm of matrices. We first derive the following result that is useful in proving Lemma 7.

Lemma 14. *Recall $\varepsilon_t = x_t'\rho + u_t$, where $\rho = (\rho_1, \rho_2, ..., \rho_K)' \in \mathbb{R}^K$ and $x_t = (\varepsilon_{t-1}, \varepsilon_{t-2}, ..., \varepsilon_{t-K})' \in \mathbb{R}^K$. Suppose that the following hold: (1) $\{u_t\}_{t=1}^T$ is an i.i.d sequence with $E(u_1^4)$ uniformly bounded. (2) the roots of $1 - \sum_{j=1}^K \rho_j L^j = 0$ are uniformly bounded away from the unit circle.*

Then we have (i) $(T-K)^{-1} \sum_{t=K+1}^T u_t^2 = O_P(1)$; (ii) $(T-K)^{-1} \sum_{t=K+1}^T x_t u_t = o_P(1)$; (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 $\{u_s : s \leq t\}$. 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 bounded. Hence, part (ii) follows by Exercise 3.77 of White (2014).

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

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

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

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

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

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

$$\hat{\varepsilon}_t = \delta_t + \varepsilon_t = \delta_t + x_t'\rho + u_t = \delta_t + (\hat{x}_t - \Delta_t)'\rho + u_t = \hat{x}_t'\rho + \tilde{u}_t. \tag{39}$$

Therefore,

$$\begin{aligned}\hat{\rho} = \left(\sum_{t=K+1}^{T}\hat{x}_t\hat{x}_t'\right)^{-1}\left(\sum_{t=K+1}^{T}\hat{x}_t\hat{\varepsilon}_t\right) &= \left(\sum_{t=K+1}^{T}\hat{x}_t\hat{x}_t'\right)^{-1}\left(\sum_{t=K+1}^{T}\hat{x}_t(\hat{x}_t'\rho + \tilde{u}_t)\right)\\ &= \rho + \left(\sum_{t=K+1}^{T}\hat{x}_t\hat{x}_t'\right)^{-1}\left(\sum_{t=K+1}^{T}\hat{x}_t\tilde{u}_t\right).\end{aligned} \tag{40}$$

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,

$$\sum_{t=K+1}^{T}\|\Delta_t\|^2 = \sum_{t=K+1}^{T}\sum_{s=t-1}^{t-K}\delta_s^2 \le K\sum_{t=1}^{T}\delta_t^2 \overset{\text{(i)}}{=} o_P(T), \tag{41}$$

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

$$\begin{aligned}\left\|\sum_{t=K+1}^{T}\left(\hat{x}_t\hat{x}_t' - x_t x_t'\right)\right\| &= \left\|\sum_{t=K+1}^{T}\left(x_t\Delta_t' + \Delta_t x_t' + \Delta_t\Delta_t'\right)\right\|\\ &\le 2\sum_{t=K+1}^{T}\|x_t\|\cdot\|\Delta_t\| + \sum_{t=K+1}^{T}\|\Delta_t\|^2\\ &\le 2\sqrt{\left(\sum_{t=K+1}^{T}\|x_t\|^2\right)\left(\sum_{t=K+1}^{T}\|\Delta_t\|^2\right)} + \sum_{t=K+1}^{T}\|\Delta_t\|^2 \overset{\text{(i)}}{=} o_P(T),\end{aligned} \tag{42}$$

where (i) follows by (41) and Lemma 14. Thus,

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

By Lemma 14, 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

$$\left[(T-K)^{-1}\sum_{t=K+1}^{T}\hat{x}_t\hat{x}_t'\right]^{-1} = O_P(1). \tag{43}$$

Step 2: show that $(T-K)^{-1}\sum_{t=K+1}^{T}\hat{x}_t\tilde{u}_t = o_P(1)$.
By Lemma 14, we have

$$(T-K)^{-1}\sum_{t=K+1}^{T} x_t u_t = o_P(1). \tag{44}$$

Notice that

$$\begin{aligned}
\left\|\frac{1}{T-K}\sum_{t=K+1}^{T}(\hat{x}_t\tilde{u}_t - x_t u_t)\right\| &= \left\|\frac{1}{T-K}\sum_{t=K+1}^{T}(\Delta_t u_t + x_t a_t + \Delta_t a_t)\right\| \\
&\le \frac{1}{T-K}\sum_{t=K+1}^{T}(\|\Delta_t u_t\| + \|x_t a_t\| + \|\Delta_t a_t\|) \\
&\le \sqrt{\left(\frac{1}{T-K}\sum_{t=K+1}^{T}\|\Delta_t\|^2\right)\left(\frac{1}{T-K}\sum_{t=K+1}^{T}u_t^2\right)} \\
&\quad + \sqrt{\left(\frac{1}{T-K}\sum_{t=K+1}^{T}\|x_t\|^2\right)\left(\frac{1}{T-K}\sum_{t=K+1}^{T}a_t^2\right)} \\
&\quad + \sqrt{\left(\frac{1}{T-K}\sum_{t=K+1}^{T}\|\Delta_t\|^2\right)\left(\frac{1}{T-K}\sum_{t=K+1}^{T}a_t^2\right)}.
\end{aligned} \tag{45}$$

We observe that

$$\begin{aligned}
\sum_{t=K+1}^{T} a_t^2 = \sum_{t=K+1}^{T}\left(\delta_t - \Delta_t'\rho\right)^2 &\le 2\sum_{t=K+1}^{T}\delta_t^2 + 2\sum_{t=K+1}^{T}(\Delta_t'\rho)^2 \\
&\le 2\sum_{t=1}^{T}\delta_t^2 + 2\|\rho\|^2\sum_{t=K+1}^{T}\|\Delta_t\|^2 \overset{\text{(i)}}{=} O_P(T),
\end{aligned} \tag{46}$$

where (i) follows by (41) and the assumption of $T^{-1}\sum_{t=1}^{T}\delta_t^2 = o_P(1)$. Combining (45) with (41) and (46), we obtain

$$\begin{aligned}
&\left\|\frac{1}{T-K}\sum_{t=K+1}^{T}(\hat{x}_t\tilde{u}_t - x_t u_t)\right\| \\
&\le \sqrt{o_P(1)\left(\frac{1}{T-K}\sum_{t=K+1}^{T}u_t^2\right)} + \sqrt{\left(\frac{1}{T-K}\sum_{t=K+1}^{T}\|x_t\|^2\right)o_P(1)} + \sqrt{o_P(1)\times o_P(1)} \overset{\text{(i)}}{=} o_P(1),
\end{aligned} \tag{47}$$

where (i) follows by Lemma 14. Now we combine (44) and (47), obtaining

$$(T-K)^{-1}\sum_{t=K+1}^{T}\hat{x}_t\tilde{u}_t = o_P(1). \tag{48}$$

By (40) together with (43) and (48), it follows that

$$\hat{\rho} - \rho = o_P(1). \tag{49}$$

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

$$\hat{u}_t - u_t = \left(\hat{\varepsilon}_t - \hat{x}_t'\hat{\rho}\right) - u_t \overset{\text{(i)}}{=} \left(\hat{x}_t'(\rho - \hat{\rho}) + \tilde{u}_t\right) - u_t = \hat{x}_t'(\rho - \hat{\rho}) + a_t, \tag{50}$$

where (i) follows by (39). Therefore, we have

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

where (i) follows by (42), (49), (46) and Lemma 14.

To see the pointwise result, we notice that by (50) and (49), 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. □

C Figures and Tables

Figure 2: Raw Data

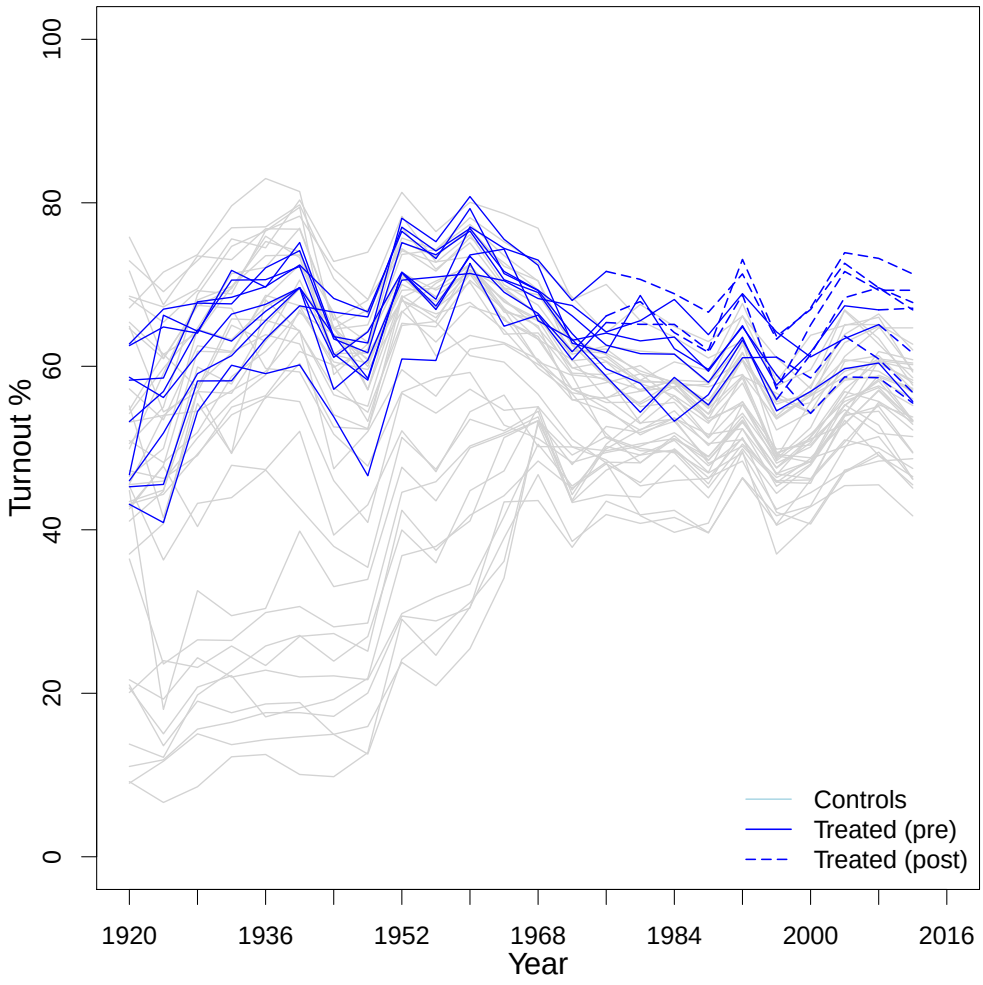


Figure 3: Maine, Minnesota, and Wisconsin

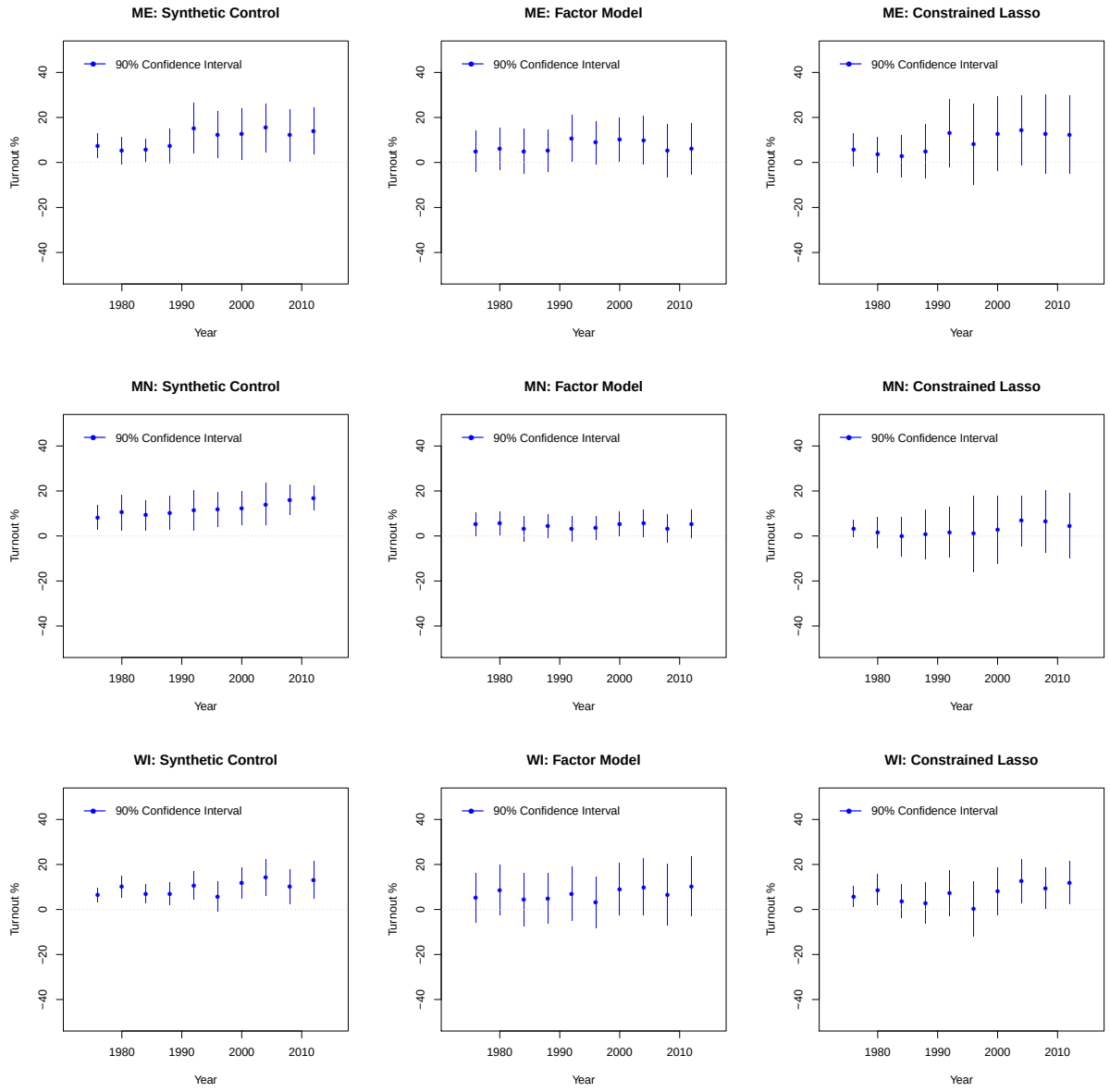


Figure 4: Idaho, New Hampshire, and Wyoming

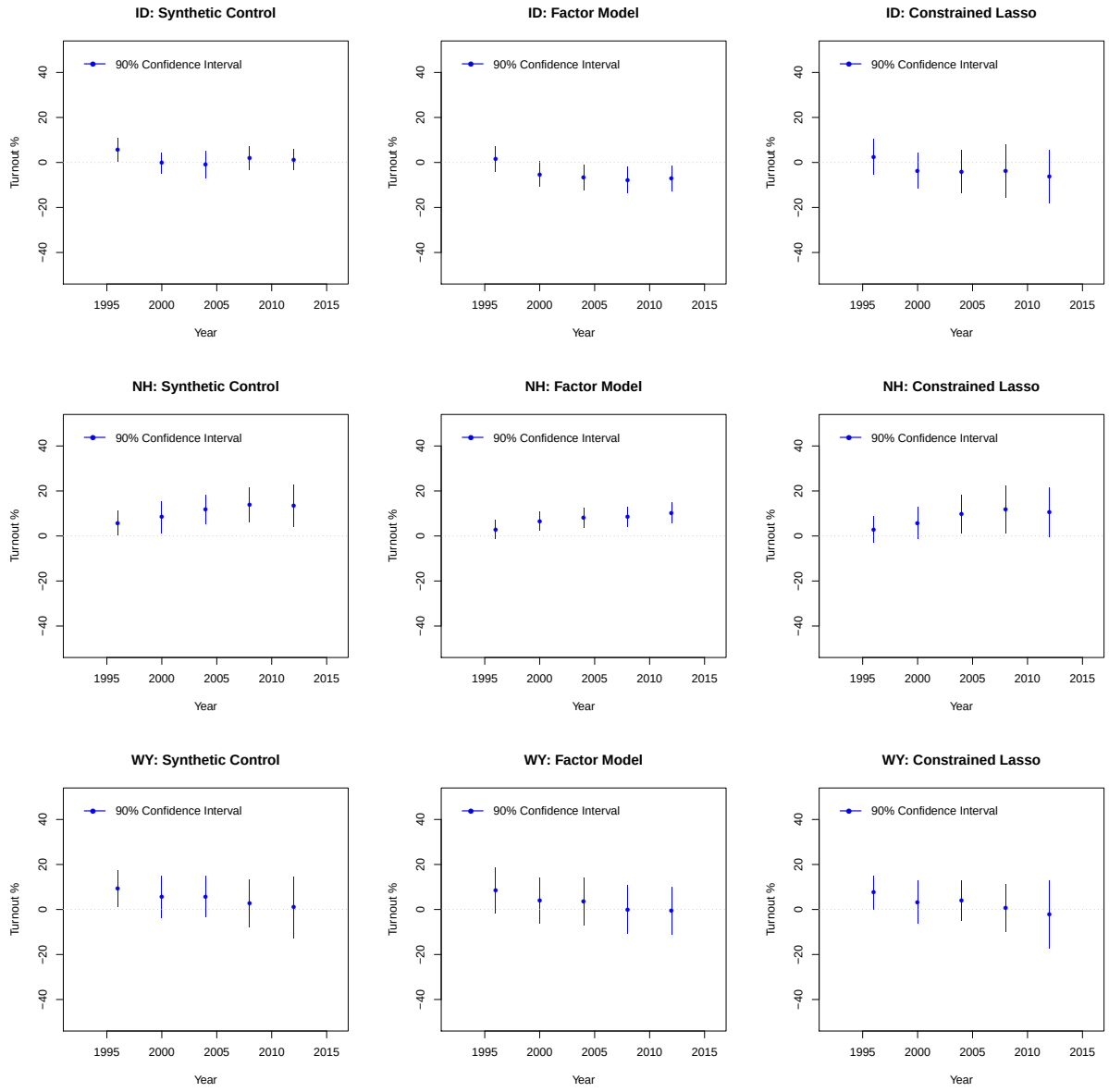


Figure 5: Montana, Iowa, Connecticut

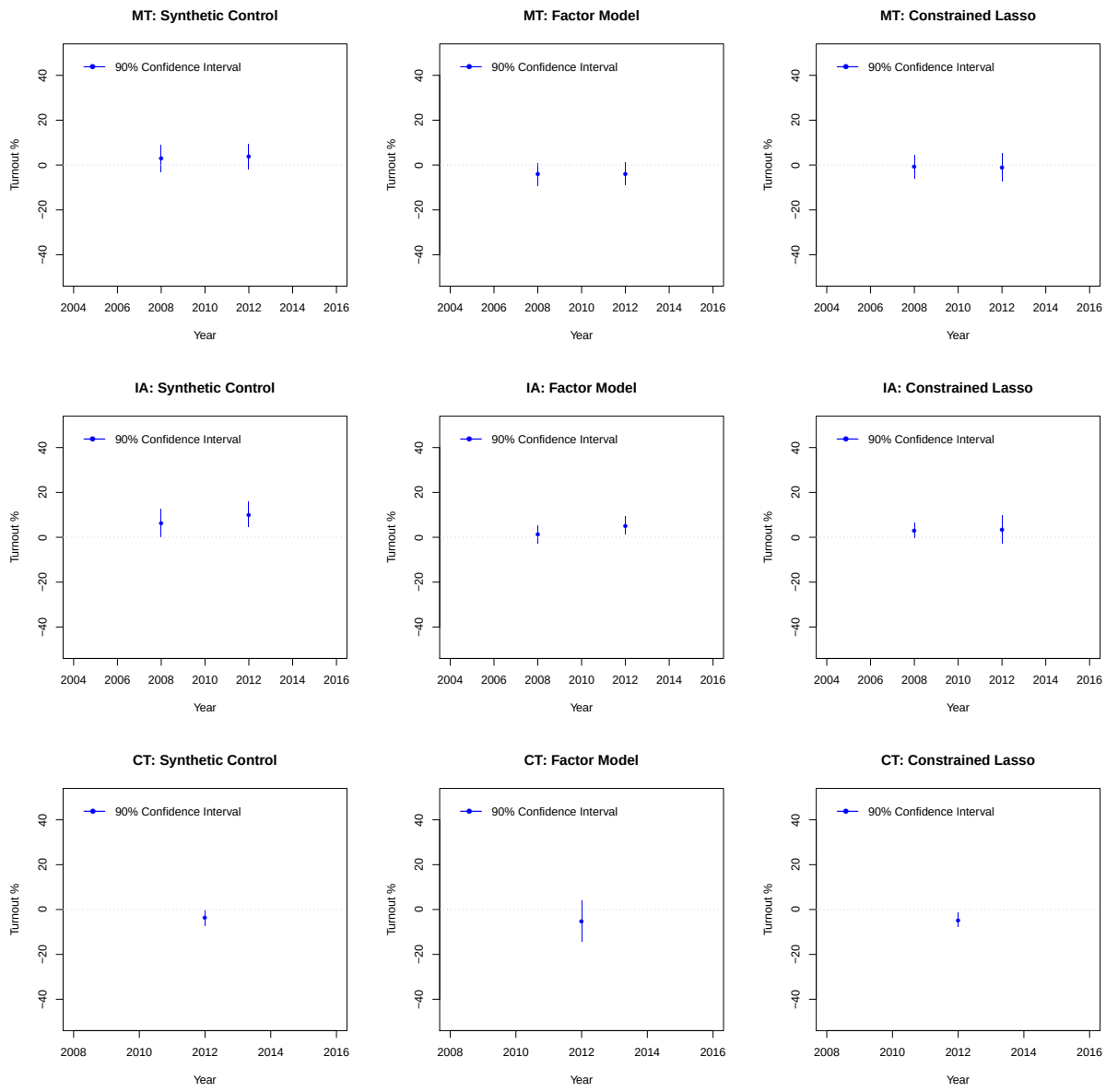


Table 1: Size DGP1

	DGP1a								
	i.i.d. data with $\rho_\epsilon = \rho_u = 0$								
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.09	0.09	0.09	0.09	0.09	0.08	0.09	0.09	0.09
$T_0 = 50$	0.10	0.10	0.09	0.10	0.09	0.10	0.09	0.10	0.10
$T_0 = 100$	0.10	0.11	0.10	0.10	0.10	0.10	0.10	0.11	0.10
	Weakly dependent data with $\rho_\epsilon = \rho_u = 0.6$								
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.10	0.11	0.11	0.12	0.12	0.12	0.11	0.11	0.11
$T_0 = 50$	0.12	0.12	0.12	0.13	0.12	0.12	0.12	0.13	0.12
$T_0 = 100$	0.12	0.11	0.11	0.13	0.11	0.10	0.13	0.11	0.11
	DGP1b								
	i.i.d. data with $\rho_\epsilon = \rho_u = 0$								
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.10	0.10	0.10	0.10	0.09	0.10	0.10	0.09	0.10
$T_0 = 50$	0.09	0.10	0.08	0.09	0.09	0.09	0.09	0.11	0.08
$T_0 = 100$	0.11	0.09	0.11	0.11	0.09	0.10	0.11	0.09	0.10
	Weakly dependent data with $\rho_\epsilon = \rho_u = 0.6$								
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.11	0.12	0.10	0.13	0.13	0.14	0.12	0.13	0.10
$T_0 = 50$	0.12	0.13	0.12	0.11	0.12	0.12	0.13	0.12	0.12
$T_0 = 100$	0.11	0.10	0.11	0.10	0.11	0.11	0.11	0.11	0.11

Notes: Simulation design is described in the main text. Nominal level = 0.1. Based on simulations with 2000 repetitions.

Table 2: Size DGP2

	DGP2a								
	i.i.d. data with $\rho_\epsilon = \rho_u = 0$								
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.09	0.09	0.10	0.09	0.09	0.10	0.10	0.09	0.10
$T_0 = 50$	0.11	0.09	0.10	0.11	0.10	0.10	0.11	0.09	0.10
$T_0 = 100$	0.09	0.11	0.10	0.09	0.10	0.10	0.09	0.11	0.11
	Weakly dependent data with $\rho_\epsilon = \rho_u = 0.6$								
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.12	0.12	0.11	0.12	0.13	0.13	0.12	0.12	0.11
$T_0 = 50$	0.13	0.12	0.12	0.13	0.12	0.12	0.13	0.14	0.13
$T_0 = 100$	0.12	0.12	0.12	0.12	0.11	0.11	0.13	0.12	0.12
	DGP2b								
	i.i.d. data with $\rho_\epsilon = \rho_u = 0$								
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.10	0.11	0.09	0.09	0.10	0.09	0.10	0.10	0.09
$T_0 = 50$	0.10	0.11	0.10	0.10	0.11	0.09	0.10	0.11	0.09
$T_0 = 100$	0.11	0.09	0.11	0.10	0.10	0.12	0.11	0.10	0.10
	Weakly dependent data with $\rho_\epsilon = \rho_u = 0.6$								
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.11	0.12	0.11	0.12	0.12	0.11	0.11	0.10	0.09
$T_0 = 50$	0.10	0.12	0.11	0.10	0.11	0.11	0.11	0.12	0.11
$T_0 = 100$	0.10	0.11	0.13	0.11	0.10	0.11	0.11	0.12	0.10

Notes: Simulation design is described in the main text. Nominal level = 0.1. Based on simulations with 2000 repetitions.

Table 3: Power DGP1

	DGP1a								
	i.i.d. data with $\rho_\epsilon = \rho_u = 0$								
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.49	0.45	0.43	0.38	0.38	0.39	0.50	0.47	0.46
$T_0 = 50$	0.57	0.55	0.55	0.50	0.47	0.49	0.57	0.57	0.56
$T_0 = 100$	0.61	0.58	0.59	0.56	0.51	0.50	0.61	0.59	0.60
	Weakly dependent data with $\rho_\epsilon = \rho_u = 0.6$								
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.56	0.55	0.55	0.49	0.50	0.52	0.59	0.58	0.57
$T_0 = 50$	0.59	0.57	0.58	0.54	0.55	0.55	0.62	0.60	0.62
$T_0 = 100$	0.62	0.60	0.60	0.59	0.58	0.55	0.64	0.62	0.63
	DGP1b								
	i.i.d. data with $\rho_\epsilon = \rho_u = 0$								
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.53	0.53	0.53	0.38	0.43	0.47	0.50	0.50	0.50
$T_0 = 50$	0.59	0.59	0.58	0.52	0.54	0.57	0.58	0.57	0.57
$T_0 = 100$	0.61	0.61	0.60	0.55	0.58	0.60	0.60	0.59	0.60
	Weakly dependent data with $\rho_\epsilon = \rho_u = 0.6$								
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.63	0.63	0.66	0.52	0.57	0.64	0.61	0.61	0.62
$T_0 = 50$	0.64	0.65	0.67	0.56	0.61	0.62	0.64	0.64	0.66
$T_0 = 100$	0.63	0.65	0.66	0.58	0.60	0.61	0.64	0.64	0.65

Notes: Simulation design is described in the main text. Nominal level = 0.1. Based on simulations with 2000 repetitions.

Table 4: Power DGP2

	DGP2a								
	i.i.d. data with $\rho_\epsilon = \rho_u = 0$								
	Synthetic control			Factor model			constr. Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.49	0.52	0.51	0.32	0.41	0.46	0.46	0.48	0.49
$T_0 = 50$	0.55	0.55	0.58	0.45	0.52	0.57	0.55	0.54	0.55
$T_0 = 100$	0.54	0.57	0.59	0.49	0.55	0.60	0.55	0.57	0.60
	Weakly dependent data with $\rho_\epsilon = \rho_u = 0.6$								
	Synthetic control			Factor model			constr. Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.60	0.62	0.64	0.49	0.57	0.63	0.58	0.61	0.62
$T_0 = 50$	0.61	0.63	0.65	0.53	0.58	0.62	0.61	0.64	0.65
$T_0 = 100$	0.61	0.63	0.67	0.54	0.59	0.63	0.61	0.63	0.67
	DGP2b								
	i.i.d. data with $\rho_\epsilon = \rho_u = 0$								
	Synthetic control			Factor model			constr. Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.16	0.17	0.19	0.21	0.28	0.37	0.37	0.38	0.38
$T_0 = 50$	0.18	0.20	0.21	0.26	0.39	0.46	0.43	0.44	0.45
$T_0 = 100$	0.18	0.22	0.25	0.28	0.40	0.51	0.44	0.47	0.50
	Weakly dependent data with $\rho_\epsilon = \rho_u = 0.6$								
	Synthetic control			Factor model			constr. Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.21	0.23	0.28	0.30	0.36	0.45	0.43	0.42	0.46
$T_0 = 50$	0.23	0.25	0.29	0.34	0.43	0.52	0.47	0.48	0.51
$T_0 = 100$	0.22	0.24	0.28	0.35	0.45	0.54	0.47	0.50	0.51

Notes: Simulation design is described in the main text. Nominal level = 0.1. Based on simulations with 2000 repetitions.

Table 5: p-Value: no effect

	Moving Block Permutations			i.i.d. Permutations		
	Synth. Control	Factor Model	Constr. Lasso	Synth. Control	Factor Model	Constr. Lasso
CT	0.08	0.29	0.04	0.08	0.29	0.04
IA	0.04	0.25	0.29	0.01	0.2	0.26
ID	0.83	0.04	0.42	0.7	0.04	0.44
ME	0.04	1	0.83	0	1	0.91
MN	0.04	0.96	0.58	0	0.93	0.54
MT	0.38	0.33	0.96	0.32	0.26	0.9
NH	0.04	0.21	0.38	0	0.09	0.33
WI	0.04	0.92	0.17	0	0.72	0.05
WY	0.46	0.25	0.62	0.42	0.37	0.65