THE UNIVERSITY OF CHICAGO

ESSAYS ON THE ECONOMETRICS OF DEPENDENT DATA

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
JIANFEI CAO

CHICAGO, ILLINOIS
JUNE 2021

To Xiaoxia, who brought me to the world of Econometrics.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am greatly indebted to my advisors Christian Hansen, Max Farrell, Tetsuya Kaji, and Panos Toulis for their invaluable insight and support. Their influence on me has been enormous, both for this dissertation and me as an econometrician.

Throughout writing this dissertation, I've been receiving help from numerous professors, colleges, and friends. I am grateful in particular to Stéphane Bonhomme, Connor Dowd, Damian Kozbur, Tengyuan Liang, Tesary Lin, Shirley Lu, Azeem Shaikh, Xiaoxia Shi, Matt Shum, Max Tabord-Meehan, Alex Torgovitsky, Ruey Tsay, Lucciano Villacorta, and Yike Wang. I thank all seminar participants at the weekly meetings of the econometrics advising group at UChicago, annual meetings of the Econometric Society, and the Midwest Econometrics Group for their constructive and encouraging feedback. I thank all my friends, colleagues, and staffs in Booth Ph.D. program for helping me navigate through this journey. I'm extremely grateful to my parents for their endless love and support. Special thanks to my cats, Ily, Pika, Simba, and Hoopa, who have been great company over the years.

ABSTRACT

This dissertation studies the estimation and statistical inference of a few methods that are commonly used in empirical studies of economics, when the data is dependent. Chapter 1 studies the linear IV models with clustering dependence, which are widely used in empirical studies. The common solution, the *cluster covariance estimator*, often produces undesirable inferential results, especially with weak instruments. I propose a method that is robust to both weak IV and (potentially heterogeneous) clustering dependence. The proposed method is based on the idea of Fama-MacBeth estimation, with group-level estimators being a truncated version of the unbiased IV estimator. Asymptotic validity is shown under both strong and weak IV sequences, as well as under general requirements. Simulation results indicate that the method has good finite-sample performance in both size and power. The proposed method is applied to study the effect of city compactness on population density.

Chapter 2, coauthored with Damian Kozbur, Christian Hansen, and Lucciano Villacorta, presents and analyzes an approach to inference for dependent data. The primary setting considered here is with spatially indexed data in which the dependence structure of observed random variables is characterized by an observed dissimilarity measure over spatial indexes. Observations are partitioned into clusters with the use of an unsupervised clustering algorithm applied to the dissimilarity measure. Once the partition into clusters is learned, a cluster-based inference procedure is applied to a statistical hypothesis test. The procedure proposed in the paper allows the number of clusters to depend on the data, which gives researchers a principled method for choosing an appropriate clustering level. The paper gives conditions under which the proposed procedure asymptotically attains the correct size.

Chapter 3, coauthored with Connor Dowd, studies the synthetic control methods in the presence of spillover effects. The synthetic control method is often used in treatment effect estimation with panel data where only a few units are treated and a small number of post-treatment periods are available. Current estimation and inference procedures for synthetic

control methods do not allow for the existence of spillover effects, which are plausible in many applications. In this chapter, we consider estimation and inference for synthetic control methods, allowing for spillover effects. We propose estimators for both direct treatment effects and spillover effects and show that they are asymptotically unbiased. In addition, we propose an inferential procedure and show that it is asymptotically unbiased. Our estimation and inference procedure applies to cases with multiple treated units and/or multiple post-treatment periods, and to ones where the underlying factor model is either stationary or cointegrated.

CHAPTER 1

ROBUST IV INFERENCE WITH CLUSTERING DEPENDENCE

1.1 Introduction

In linear IV models, accounting for clustering dependence has been a standard procedure in conducting statistical inference in empirical research. A common solution is to use the cluster covariance estimator (CCE), which is often referred to as the "clustered standard error" method. In linear models CCE methods are shown to deliver valid inference under either strong homogeneity across groups (the *large-homogeneous-group* approach, e.g., Bester et al., 2011a) or lots of small groups (the *many-small-group* approach, e.g., Hansen and Lee, 2019). Those results provide theoretical justification for the usage of CCE methods in the linear IV model under strong IV.[1] This chapter concerns statistical inference in the linear IV model with clustering dependence.

However, for many common settings, whether either the large-homogeneous-group or many-small-group approach can justify the usage of standard clustering methods is not clear. Specifically, Bester et al. (2011a) assume all groups are similar in size and the same design-matrix limit, which does not hold in many settings such as clustering by state. Hansen and Lee (2019) require that $\max_g n_g^2/n \to 0$, where n_g is the number of observations in group g and n is the sample size. In the case of equal-sized groups, this requirement implies $n/G^2 \to 0$, where G is the number of groups. MacKinnon and Webb (2017) conduct simulation studies and show that the empirical rejection can be as high as 0.1073 at a level-0.05 test, when $(n, G) = (2000, 50)$, thus with $n/G^2 = 0.8$, and group sizes are proportional to population of the 50 states in the US. A non-exhaustive search in recent empirical research suggests

1. Hansen and Lee (2019) cover both OLS and IV, whereas Hansen (2007) considers only the OLS case, but the results can be extended to the strong IV model.

n/G^2 is often large, for example, Coibion et al. (2017) with $n/G^2 = 0.46$ or 0.34 in Table 3, Dell (2012) with n/G^2 ranging from 1.21 to 16.65 in Table 7, and Deryugina et al. (2019) with $n/G^2 = 2.43$ in Table 2. For a discussion on the poor asymptotic approximation of inference methods based on asymptotic theory, see Ferman and Pinto (2019a), Ferman (2019), MacKinnon and Webb (2017), and Young (2019).

Moreover, standard methods suffer from size distortion when weak IV is a concern. Although robust inference methods such as the Anderson-Rubin test (AR, Anderson and Rubin, 1949) work under standard assumptions described in the previous paragraph, whether those methods have good inferential properties when standard assumptions break is not well understood. In the simulation section, I show that the extension of AR with the standard error calculated by CCE methods can result in size distortion under imbalanced group sizes, with sizes being as high as 0.116 at a level-0.05 test.

Alternatively, Fama-MacBeth methods (Fama and MacBeth, 1973b; Ibragimov and Müller, 2010), sometimes referred to as mean group estimation (Pesaran and Smith, 1995; Pesaran et al., 1999), provide another inferential approach that exploits the clustering dependence structure. Those methods first perform group-level estimation for each group and consider a weighted average of all group-level estimators. Under a wide variety of circumstances[2], the resulting average has well-understood properties, and a simple procedure such as a t-test can be used to attain valid inference. In this chapter I introduce a group-based inference method that is built on Fama-MacBeth methods, in order to simultaneously solve clustering dependence and potentially weak IV.

In this chapter I study robust inferential methods to overcome the practical issues mentioned above, based on the idea of Fama-MacBeth estimation. Because the Fama-MacBeth approach calculates the group-level estimator using only the data in a certain group, a potential finite-sample problem may arise in the IV estimation. To account for that possibility,

2. For example, Ibragimov and Müller (2010) only require the group-level estimators to be asymptotically normal, without homogeneity across groups.

I propose a truncated version of the unbiased IV estimator introduced by Andrews and Armstrong (2017) in calculating the group-level estimator. I show that this estimator is nearly unbiased, and that using it in the Fama-MacBeth approach produces valid inference. The proposed method allows for a moderate number of moderate-sized groups (e.g., 30 groups of around 30 observations as in the simulation section) and is robust to both weak IV and heterogeneous clustering dependence. Table 1.1 summarizes whether a certain aforementioned method is robust to a non-conventional set-up.

Table 1.1: Robustness of Inferential Methods in Linear IV Models with Clustering Dependence

	$n/G^2 \gg 0$	Heterogeneous Groups	Weak IV
CCE (large-G)	NO	YES	NO
CCE (small-G)	YES	NO	NO
AR-CCE (large-G)	NO	YES	YES
AR-CCE (small-G)	YES	NO	YES
Fama-MacBeth	YES	YES	NO
Proposed method	YES	YES	YES

Notes: This table roughly summarizes whether a candidate inferential method is robust to a certain non-conventional set-up. "YES" means it generally delivers correct size and "NO" means it does not. "Large-G" stands for the *many-small-group* approach and "small-G" stands for the *large-homogeneous-group*" one. "AR-CCE" is the natural extension of the Anderson-Rubin method to the case with clustering dependence (described in Section 1.4.2). "Proposed method" is the Fama-MacBeth approach with truncated unbiased estimators proposed in this chapter.

Both an unbiased group-level estimator and the truncation are important in implementing the Fama-MacBeth approach in this setting. Without the former, the group-level IV estimator may lead to substantial finite-sample bias and cause size distortion under the null. The latter guarantees the group-level estimators have finite second moments such that the test has power. Simulation studies show that direct usage of Andrews and Armstrong (2017)

produces far less power, and the proposed method is robust to many settings and has good power properties.

Throughout, I assume one endogenous variable and focus on the case of one instrument. Cases with multiple instruments can be dealt with using the averaging method introduced by Andrews and Armstrong (2017). Similar to Andrews and Armstrong (2017), to implement the proposed method, the sign of the first-stage parameter is assumed to be known. This assumption is often a weak one in empirical studies, because the sign of the instrument is typically embedded in the reasoning of instrument validity and comes in before the discussion of the strength of the instrument. For instance, Mills (2019) shows that 82.35% of the papers published in the American Economic Review from 2014 to 2018 and with "instrument" in the abstract claim the first-stage sign is known. Additionally, Mills (2019) shows that exploiting information of the first-stage sign may help improve test power. Another underlying assumption throughout is the group-level normal model (see Section 1.3.1), for which a sufficient assumption would be weak dependence as in the large-homogeneous-group approach (Bester et al., 2011a).

The chapter contributes to two streams of literature. First, the proposed method fills a gap in the literature on cluster-based inferential methods. Although those methods are extensively studied under standard assumptions such as the large-homogeneous-group case and the many-small-group case (Bertrand et al., 2004; Hansen, 2007; Bester et al., 2011a; Cameron and Miller, 2015; Hansen and Lee, 2019), the properties of those methods outside the standard assumptions are largely unknown. I show through simulation that existing methods can break under many circumstances. I advocate the usage of the proposed Fama-MacBeth approach with truncated unbiased IV estimation and show its validity.

Second, this chapter complements the recent literature on the Fama-MacBeth approach and shows its usefulness. Fama and MacBeth (1973b) introduced this approach, but it was only recently theoretically justified by Ibragimov and Müller (2010). Ibragimov and Müller

(2010), Canay et al. (2017), Cao et al. (2019), and Hagemann (2019a,b) have documented the robustness and good power properties of this approach. Many of their results can be either applied or extended to the strong IV case, but extension to allowing for weak IV is non-trivial.

The remainder of the chapter is organized as follows. Section 1.2 introduces a truncated version of the unbiased IV estimator with known first-stage sign. Section 1.3 proposes the inferential method that applies the truncated unbiased IV estimator. In addition, the primitive conditions for both strong and weak IV asymptotics are listed. Simulation studies are presented in Section 1.4. In Section 1.5, I apply the proposed method to study the effect of city compactness on population density. Section 1.6 concludes. Proofs are relegated to the appendix.

1.2 Truncated Unbiased IV with Known First-Stage Sign

We first consider a simple linear IV model. Let X, Y, and Z be $n \times 1$ data vectors for three scalar variables. The reduced-form formulation of the linear IV model is

$$\begin{cases} Y = Z\pi\beta + U, \\ X = Z\pi + V, \end{cases} \tag{1.1}$$

where π and β are both scalars. We are interested in the structural equation parameter β. Assume the sign of π is known, and, without loss of generality, let $\pi > 0$. Assume the vector of reduced-form and first-stage estimators follows

$$\widehat{\psi} = \begin{pmatrix} \widehat{\gamma} \\ \widehat{\pi} \end{pmatrix} = \begin{pmatrix} (Z'Z)^{-1}Z'Y \\ (Z'Z)^{-1}Z'X \end{pmatrix} \sim N(\mu, \Sigma), \tag{1.2}$$

where

$$\mu = \begin{pmatrix} \pi\beta \\ \pi \end{pmatrix}, \ \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

The usual IV estimator is $\widehat{\beta}_{IV} = \widehat{\gamma}/\widehat{\pi}$. We assume throughout that Σ is known and positive definite. Our analysis relies heavily on (1.2), which applies to cases where normality is a good approximation of the reduced-form and first-stage coefficients $\widehat{\psi}$.

Remark 1.1. *The model* (1.1) *has an equivalent structural formulation. Generalization of* (1.1) *to models with multiple instruments and/or other control variables can be done in standard methods. For multiple instruments, we can use a weighted average of the proposed estimators of a single instrument, because a weighted average of (nearly) unbiased estimators is still (nearly) unbiased. Including other control variables can be done through projection on the null space by the Frisch-Waugh-Lovell theorem.*

Remark 1.2. *The assumption that π has known sign is often weak in empirical studies. According to a survey by Mills (2019) on papers published in the American Economic Review from 2014 to 2018, 14 out of 17 papers with "instrument" in the abstract claim to have known first-stage sign.*

Remark 1.3. *The normal model* (1.2) *is common in the literature on IV inference that is robust to weak instruments (see, e.g., Andrews et al., 2006; Andrews and Mikusheva, 2016; Kleibergen, 2002; Moreira, 2003; Moreira and Moreira, 2019; Staiger and Stock, 1997). One motivation is that the vector $(\pi\beta, \beta)$ can be considered a regular parameter and well estimated under mild regularity conditions, whereas β itself is only weakly regular in the case of weak instruments (Kaji, 2021). As a result, the least-squares estimator for $(\pi\beta, \beta)$ can often be approximated by a normal distribution. One simple example for the model* (1.2) *to hold is the case where Z is fixed and the rows of $[U, V]$ are i.i.d. or stationary. In this case, the*

covariance matrix of $\widehat{\psi}$ *is*

$$\Sigma = (I_2 \otimes (Z'Z)^{-1}Z')\text{Var}[(U', V')'](I_2 \otimes (Z'Z)^{-1}Z')' \tag{1.3}$$

and can be consistently estimated. See Andrews et al. (2019) for a review on the normal approximation to the distribution of $(\widehat{\gamma}, \widehat{\pi})$.

We follow Andrews and Armstrong (2017) and define the unbiased IV estimator. Let

$$\widehat{\delta} = \widehat{\delta}(\widehat{\psi}, \Sigma) = \widehat{\gamma} - \frac{\sigma_{12}}{\sigma_2^2}\widehat{\pi}$$

and

$$\widehat{\tau} = \widehat{\tau}(\widehat{\psi}, \Sigma) = \frac{1}{\sigma_2}\frac{1 - \Phi(\widehat{\pi}/\sigma_2)}{\phi(\widehat{\pi}/\sigma_2)} = \frac{1}{\sigma_2}\Psi(\widehat{\pi}/\sigma_2),$$

where $\Psi(x) = (1 - \Phi(x))/\phi(x)$, and $\Phi(\cdot)$ and $\phi(\cdot)$ are cdf and pdf for the standard normal distribution, respectively. The unbiased IV estimator is

$$\widehat{\beta}_U = \widehat{\beta}_U(\widehat{\psi}, \Sigma) = \widehat{\delta}\widehat{\tau} + \frac{\sigma_{12}}{\sigma_2^2}.$$

It is shown that $E[\widehat{\beta}_U] = \beta$ when $\pi > 0$.

Remark 1.4. *The main idea of* $\widehat{\beta}_U$ *is to use the fact that* $\widehat{\tau}$ *is an unbiased estimator for* $1/\pi$ *(Voinov and Nikulin, 1993). Because* $\widehat{\delta}$ *can be considered the projection of* $\widehat{\gamma}$ *on the null space of* $\widehat{\pi}$, $\widehat{\delta}$ *is independent of* $\widehat{\pi}$, *and thus of* $\widehat{\tau}$ *as a function of* $\widehat{\pi}$. *Those facts lead to* $E[\widehat{\beta}_U] = \beta$ *(Andrews and Armstrong, 2017).*

Define the truncated version of the unbiased IV estimator by

$$\tilde{\beta} = \widehat{\delta}\tilde{\tau} + \frac{\sigma_{12}}{\sigma_2^2},$$

where

$$\tilde{\tau} = \frac{1}{\sigma_2}\Psi\left(\frac{\max\{\widehat{\pi}, \pi^*\}}{\sigma_2}\right),$$

and π^* is some truncation parameter. That is, we "winsorize" the unbiased IV estimator according to $\widehat{\pi}$ by the threshold π, when $\widehat{\pi}$ is too small. We do so because $\Psi(\cdot)$ is positive and strictly decreasing on $\mathbb{R}$, and $\Psi(x) \to \infty$ as $x \to -\infty$, which causes $\widehat{\beta}_U$ to have an unbounded second moment. By truncation, we eliminate extreme values of $\widehat{\beta}_U$, which is important in conducting inference.

Example 1.1. *We visualize the truncation in Figure 1.1. Consider a simple case where* $\widehat{\psi} = (\widehat{\gamma}, \widehat{\pi})' \sim N(\psi, I_2)$. *Then, the unbiased IV estimator for* β *is* $\widehat{\beta}_U = \widehat{\delta}\widehat{\tau}$, *where* $\widehat{\delta} = \widehat{\gamma}$ *and* $\widehat{\tau} = (1 - \Phi(\widehat{\pi}))/\phi(\widehat{\pi})$. *Define* $\widehat{\pi}_U = 1/\widehat{\tau}$, *then* $\widehat{\beta}_U = \widehat{\gamma}/\widehat{\pi}_U$; *that is,* $\widehat{\beta}_U$ *is the slope of the line through* $(\widehat{\pi}_U, \widehat{\delta})$ *and the origin. Then, the proposed truncated estimator* $\tilde{\beta}$ *is the slope of the line through* $(\tilde{\pi}, \widehat{\gamma}) = (\max\{\widehat{\pi}_U, \pi^*\}, \widehat{\delta})$ *and the origin.*

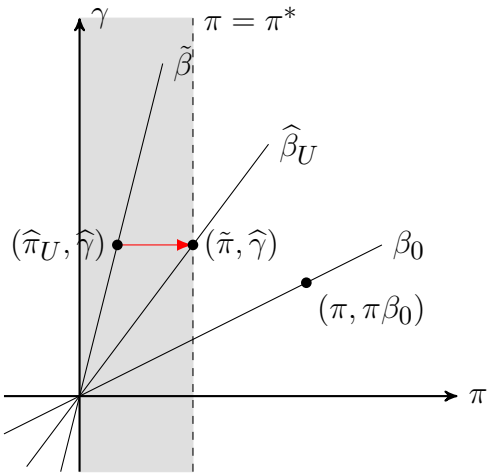


Figure 1.1: Truncation is obtained through winsorizing $\widehat{\pi}_U$ according to π^*.

The following result shows that the truncated estimator is nearly unbiased when the truncation is appropriate.

Proposition 1.1. *Assume* β *is fixed. Suppose (i)* $|\sigma_{12}/\sigma_2^2| < \infty$, *(ii)* $\pi^*/\sigma_2 \to -\infty$, *and (iii)* $\pi\pi^*/\sigma_2^2 \to -\infty$. *Then,* $E[\tilde{\beta}] - \beta \to 0$.

Remark 1.5. *Proposition 1.1 gives guidance on when the proposed estimator $\tilde{\beta}$ is approximately unbiased. A trivial example is where $\pi^* \to -\infty$ and everything else is constant, in which case, $\tilde{\beta}$ is approaching the unbiased estimator $\widehat{\beta}_U$. Under either the common strong IV asymptotics where $\sigma_2 = O(1/\sqrt{n})$ and π is constant, or the common weak IV asymptotics where $\sigma_2 = O(1/\sqrt{n})$ and $\pi = O(1/\sqrt{n})$, (ii) and (iii) require π^* to be negative and not to shrink as fast as $1/\sqrt{n}$.*

1.3 Fama-MacBeth Inference with Truncated Unbiased IV

Consider a triangular array $\{\{(X_{n,i}, Y_{n,i}, Z_{n,i})\}_{i=1}^n\}_{n\geq 1}$ that follows the linear IV model (1.1),

$$\begin{cases} Y_{n,i} = Z_{n,i}\pi_n\beta + U_{n,i}, \\ X_{n,i} = Z_{n,i}\pi_n + V_{n,i}, \end{cases} \tag{1.4}$$

and a sequence of clustering dependence structures $\{\mathcal{C}_n\}_{n\geq 1}$ with $\mathcal{C}_n = \{I_{n,g}\}_{g=1}^{G_n}$ such that $G_n \to \infty$ as $n \to \infty$, where $\mathcal{C}$ is a partition of $\{1, \ldots, n\}$. That is, for any fixed n, observations are independent across groups but may be dependent within a group. As in Section 1.2, (X, Y, Z) is considered fixed and (U, V) is considered random. The parameter of interest is β, which does not vary with the sample size n. Our goal is to make inferential statement on the hypothesis $H_0 : \beta = \beta_0$. The first-stage coefficient π_n is allowed to change with n but stays the same across groups for each fixed n.[3] In the following presentation, we suppress n for simplicity. All variables and parameters (except β) should be considered a function of n.

3. This assumption is made here for simplicity. In principle, we do not need to assume π is the same across different groups, because of the nature of group-level estimation.

1.3.1 General results

We consider a Fama-MacBeth-type procedure. Namely, we estimate a truncated unbiased IV estimator $\tilde{\beta}_g$ for each group $g \in \{1, \ldots, G\}$, using only $\{(X_i, Y_i, Z_i)\}_{i \in I_g}$. Thus, we obtain a set $\{\tilde{\beta}_g\}_{g=1}^G$ of nearly unbiased IV estimators with bounded second moments. Define group-level quantities $\{n_g, \widehat{\psi}_g, \widehat{\delta}_g, \widehat{\tau}_g, \pi_g^*\}_{g=1}^G$ accordingly. As in Section 1.2, we assume the group-level reduced-form and first-stage coefficients follow a normal distribution with known covariance Σ_g such that

$$\widehat{\psi}_g = \begin{pmatrix} \widehat{\gamma}_g \\ \widehat{\pi}_g \end{pmatrix} \sim N(\mu_g, \Sigma_g),$$

from which the group-level truncated unbiased IV estimator $\tilde{\beta}_g$ is constructed.[4] Therefore, either the errors (U, V) follow normal distribution or at least a moderate number of observations are in each group. Also, define $\{\sigma_{1,g}, \sigma_{2,g}, \sigma_{12,g}, \mu_{\delta,g}, \sigma_{\delta,g}\}$ such that

$$\begin{aligned} \Sigma_g &= \begin{pmatrix} \sigma_{1,g}^2 & \sigma_{12,g} \\ \sigma_{12,g} & \sigma_{2,g}^2 \end{pmatrix}, \\ \mu_{\delta,g} &= \pi(\beta - \sigma_{12,g}/\sigma_{2,g}^2), \\ \sigma_{\delta,g}^2 &= \sigma_{1,g}^2 - \sigma_{12,g}^2/\sigma_{2,g}^2. \end{aligned}$$

For the set of group-level estimates $\{\tilde{\beta}_g\}_{g=1}^G$, define the Fama-MacBeth estimator

$$\bar{\beta} = \frac{1}{G} \sum_{g=1}^{G} \tilde{\beta}_g$$

and the standard error

$$\text{se} = \sqrt{\frac{1}{G(G-1)} \sum_{g=1}^{G} (\tilde{\beta}_g - \bar{\beta})^2}.$$

4. In practice, $\{\Sigma_g\}_{g=1}^G$ can be estimated by model-based or HAC-type estimators.

The corresponding t-statistic is

$$t = \frac{\bar{\beta} - \beta_0}{\text{se}}.$$

We show that t is asymptotically normal when the estimator is properly truncated.

Assumption 1.1. *(i)* $\limsup_n \sup_g |\sigma_{12,g}/\sigma^2_{2,g}| < \infty$*;*

(ii) $\sup_g \pi^*_g/\sigma_{2,g} \to -\infty$*, as* $n \to \infty$*;*

(iii) $\sup_g \pi\pi^*_g/\sigma^2_{2,g} \to -\infty$*, as* $n \to \infty$*.*

Define $M = \sup_g \Psi(\pi^*_g/\sigma_{2,g})/\sigma_{2,g}$. Conceptually, M guides the overall level of truncation across groups with respect to $\widehat{\tau}_g$. The reason is that $\Psi(\cdot)$ is a strictly decreasing one-to-one map such that $\widehat{\pi}_g \geq \pi^*_g$ if and only if $\widehat{\tau}_g \leq \Psi(\pi^*_g/\sigma_{2,g})/\sigma_{2,g}$.

Assumption 1.2. *The truncation parameter* M *satisfies*

$$M = o\left(\frac{B}{\bar{\sigma}_\delta(\kappa G)^{1/3}}\right),$$

where

$$B^2 = \sum_{g=1}^{G} E[(\tilde{\beta}_g - E[\tilde{\beta}_g])^2],$$

$$\bar{\sigma}_\delta = \max_g \sigma_{\delta,g},$$

$$\kappa = \max_g K\left(-\frac{3}{2}, \frac{1}{2}; -\frac{\mu^2_{\delta,g}}{2\sigma^2_{\delta,g}}\right)$$

and $K(a, b; z)$ *is Kummer's confluent hypergeometric function.*

Remark 1.6. *Assumptions 1.1 and 1.2 are high-level conditions that allow for many IV configurations. Both a fixed* π *(strong IV) or a local drifting sequence that shrinks at the rate of* $n^{-1/2}$ *(weak IV) are discussed below. Assumption 1.1 is generally weak. Part 1.1(i) implies* σ_{12} *and* $\sigma^2_{2,g}$ *are approximately of the same scale. This assumption is reasonable*

because they are typically $O(1/n_g)$ with weak dependence. Parts 1.1(ii) & (iii) require both $\pi_g^*/\sigma_{2,g}$ and $\pi\pi_g^*/\sigma_{2,g}^2$ to go to $-\infty$, uniformly. Those assumptions are weak under strong IV as long as π_g^* is negative and bounded away from zero. Under weak IV where $\pi = O(1/\sqrt{n})$, 1.1(ii) is weak and 1.1(iii) holds when $\inf_g n_g/\sqrt{n}$ does not go to zero too fast; that is, the number of groups increases too fast. Assumption 1.2 puts restrictions on the truncation parameter. Practical suggestions of how the truncation parameters are chosen are given in Appendix A.1.

Theorem 1.1. *Under Assumption 1.1 and 1.2, $t \xrightarrow{d} N(0,1)$.*

Remark 1.7. *This result implies that the test $\psi = \mathbb{1}\{|t| > z_{\alpha/2}\}$ delivers an asymptotically correct size at level α, where $z_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. In practice, some quantities in constructing the t-statistic need to be estimated. The implementation details are in Appendix A.1.*

1.3.2 Strong IV asymptotics

In this subsection, I give the primitive assumptions under which the proposed method delivers valid inference under strong IV.

Define

$$\begin{cases} \bar{\sigma}_2 = \max_g \sigma_{2,g} \\ \underline{\sigma}_2 = \min_g \sigma_{2,g} \end{cases}. \tag{1.5}$$

Assumption S1. *(i)* $\liminf_n \pi > 0$*;*

(ii) $\limsup_n \sup_g |\sigma_{12,g}/\sigma_{2,g}^2| < \infty$*;*

(iii) $\underline{\sigma}_2 M \to \infty$ *and* $\bar{\sigma}_2 = O(1)$.

Assumption S2. *(i)* $\bar{\sigma}_2/\underline{\sigma}_2 = O(1)$*;*

(ii) $M = o(BG^{-1/3})$.

Remark 1.8. *S1(i) implies a strong IV sequence and includes the case where π is fixed as $n \to \infty$. S1(ii) is the same as Assumption 1.1(i). The first half of S1(iii) together with S2(ii) provides guidance on the choice of M. The second half of S1(iii) is weak as long as groups are not diminishing. S2(i) requires that no severe size imbalance exists across groups. Together with the assumptions under the weak IV asymptotics in Section 1.3.3, these assumptions have implications on the selection of the truncation parameter. Practical suggestions are given in Appendix A.1.*

Proposition 1.2. *Under Assumption S1 and S2 (strong IV sequence), Assumptions 1.1 and 1.2 hold.*

1.3.3 Weak IV asymptotics

In this subsection, I give the primitive assumptions under which the proposed method delivers valid inference under weak IV, where the first-stage strength parameter π follows a drifting sequence towards 0 at the rate of $n^{-1/2}$.

Let $\bar{\sigma}_2$ and $\underline{\sigma}_2$ be defined in equation (1.5). Similarly, define $\bar{\sigma}_\delta = \max_g \sigma_{\delta,g}$ and $\underline{\sigma}_\delta = \min_g \sigma_{\delta,g}$.

Assumption W1. *(i)* $\pi = \pi_0/\sqrt{n}$*;*

(ii) $\sup_n \sup_g |\sigma_{12,g}/\sigma_{2,g}^2| < \infty$*;*

(iii) $n^{-1/2}\Psi^{-1}(\underline{\sigma}_2 M)/\bar{\sigma}_2 \to -\infty$.

Assumption W2. *(i)* $\pi^2/\underline{\sigma}_\delta^2 \to 0$*;*

(ii) $M = o(B\bar{\sigma}_\delta^{-1}G^{-1/3})$.

Remark 1.9. *W1(i) is standard in the weak IV literature (e.g., Staiger and Stock, 1997). In the case of weak dependence with approximately balanced groups, $\sigma_{2,g} = O(n_g^{-1/2})$, so W1(iii) implies $\Psi^{-1}(\underline{\sigma}_2 M)/\sqrt{G} \to -\infty$; $\underline{\sigma}_\delta = O(1/\min_g n_g)$, so W2(i) implies $\max_g n_g/n \to 0$ (cf. $\max_g n_g^2/n \to 0$ in Hansen and Lee, 2019).*

Proposition 1.3. *Under Assumptions W1 and W2 (weak IV sequence), Assumptions 1.1 and 1.2 hold.*

1.4 Simulation

In this section, we study the finite-sample performance of the proposed estimator. In all the following settings, the data generating process follows the linear IV model (1.4), where $n = 900$ and $G = 30$ such that $n/G^2 = 1$, which deviates from the usual asymptotics. The null hypothesis is $H_0 : \beta = 0$. For each setting, 1,000 replications are conducted to calculate the empirical rejection rate.

For each setting, we observe $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ and a partition $\{I_g\}_{g=1}^G$ of $\{i\}_{i=1}^n$. Let consecutive observations belong to the same group; that is, $I_1 = \{1, 2, \ldots, |I_1|\}$, $I_2 = \{|I_1| + 1, \ldots, |I_1| + |I_2|\}$, and so on, where $|\cdot|$ is cardinality. The data are drawn according to the following process:

$$
\begin{aligned}
&Y_i = Z_i\pi\beta + U_i \\
&X_i = Z_i\pi + V_i \\
&\begin{pmatrix} U_i \\ V_i \end{pmatrix} \sim N\left(0, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right), \text{ if } i = 1 + \sum_{h=1}^{g} |I_h| \text{ for some } g = 0, 1, \ldots, G-1 \\
&\begin{pmatrix} U_i \\ V_i \end{pmatrix} = 0.5 \begin{pmatrix} U_{i-1} \\ V_{i-1} \end{pmatrix} + \sqrt{1 - 0.5^2} \begin{pmatrix} \varepsilon_i^U \\ \varepsilon_i^V \end{pmatrix}, \text{ if } i \neq 1 + \sum_{h=1}^{g} |I_h| \text{ for any } g = 0, 1, \ldots, G-1 \\
&\begin{pmatrix} \varepsilon_i^U \\ \varepsilon_i^V \end{pmatrix} \sim N\left(0, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right) \text{ and is i.i.d. across } i.
\end{aligned}
$$

Also, each dimension of the k-dimensional instruments Z_i takes one draw from the distribution of $\{U_i\}_{i=1}^n$ and is fixed across replications. Thus, (U_i, V_i) within each group follows an AR(1) process and is independent across different groups. The parameters $(\beta, \pi, \{I_g\}_{g=1}^G, k)$

vary accordingly across settings.

1.4.1 Debiasing and truncation

We first investigate three Fama-MacBeth-type inferential procedures and show the necessity of debiasing and truncation. We consider the t-test on group-level 2SLS estimators (FM), the t-test on group-level unbiased IV estimators(FMU), and the proposed t-test on group-level truncated unbiased IV estimators (FMUT), with truncation parameter selected as suggested in Appendix A.1. The full-sample 2SLS with CCE estimates of standard errors is also reported for comparison.

In this experiment, we have five instrumental ($k = 5$) and one endogenous variable. Groups are imbalanced in sizes, with five groups of 90 observations and 25 groups of 18 observations. For each group, the observations follow an AR(1) process as described before. The first-stage coefficient is $\pi = (0.1, 0.1, 0.1, 0.1, 0.1)'/\sqrt{5}$ such that $\|\pi\|_2 = 0.1$.

The power curves are reported in Figure 1.2. Estimators used in CCE and FM are both biased. FM has large bias between the two, because it uses group-level 2SLS estimators with much larger finite-sample bias than the full-sample estimator. FMU is less powerful with FMUT, because the unbiased IV estimator does not have a bounded second moment, such that the resulting t-statistic has a tail that is too large.

1.4.2 Comparison with other methods

Here, we compare the proposed method with existing inferential procedures. We consider the "clustered standard error" approach (CCE) and the natural extension of Anderson-Rubin test to our settings (AR-CCE). To implement the AR-CCE method, we apply CCE to the regression of $Y - X\beta_0$ on Z, where β_0 is the hypothesized value as in $H_0 : \beta = \beta_0$. In our case, we test $H_0 : \beta = 0$, so AR-CCE is equivalent to performing CCE to test the hypothesis $H_0 : \gamma = 0$ in the regression $Y = Z\gamma + U$.

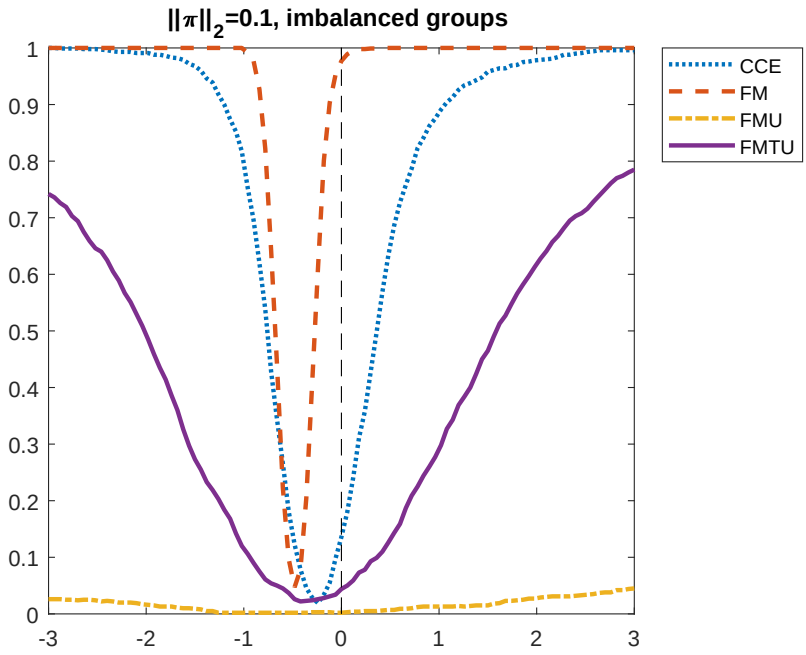


Figure 1.2: Power comparison among Fama-MacBeth procedures ($\alpha = 0.05$)

We look at several configurations. The number of instruments k varies in the set $\{1, 5, 10\}$. The first-stage strength is chosen such that $\|\pi\|_2 \in \{0.1, 0.5\}$, with $\pi = \|\pi\|_2 \iota_k / \sqrt{k}$ and ι_k being a k-vector of 1's. For example, in the case of $\|\pi\|_2 = 0.1$ and $k = 5$, we have $\pi = (0.1, 0.1, 0.1, 0.1, 0.1)'/\sqrt{5}$. We also consider both balanced and imbalanced groups. In the balanced-group case, we have 30 groups of 30 observations; in the imbalanced-group case, we have 5 groups of 90 observations and 25 groups of 18 observations.

The sizes are reported in Table 1.2 and the power curves are in Figures 1.3, 1.4, and 1.5. Among all methods, only FMUT is able to deliver a robust inference result at the null across all settings. CCE displays a noticeable bias under $\|\pi\|_2$ and over-identification. AR-CCE is robust to weak instruments, but not under group imbalance.

Table 1.2: Simulation results: comparison on estimation and inference ($\alpha = 0.05$)

			Balanced Groups			Imbalanced Groups		
			Median	MAD	Size	Median	MAD	Size
$k = 1$	$\pi = 0.5$	CCE	0.002	0.050	0.040	0.001	0.049	0.062
		AR-CCE	-	-	0.040	-	-	0.060
		FMTU	-0.010	0.057	0.039	-0.035	0.087	0.052
	$\pi = 0.1$	CCE	0.010	0.254	0.043	0.007	0.251	0.048
		AR-CCE	-	-	0.040	-	-	0.060
		FMTU	0.032	0.303	0.066	0.002	0.354	0.048
$k = 5$	$\pi = 0.5$	CCE	0.011	0.051	0.051	0.010	0.052	0.063
		AR-CCE	-	-	0.037	-	-	0.092
		FMTU	-0.068	0.110	0.033	-0.078	0.141	0.034
	$\pi = 0.1$	CCE	0.194	0.256	0.119	0.202	0.256	0.136
		AR-CCE	-	-	0.037	-	-	0.092
		FMTU	0.073	0.260	0.047	0.029	0.312	0.044
$k = 10$	$\pi = 0.5$	CCE	0.022	0.051	0.076	0.021	0.050	0.082
		AR-CCE	-	-	0.063	-	-	0.116
		FMTU	-0.055	0.105	0.046	-0.074	0.131	0.037
	$\pi = 0.1$	CCE	0.277	0.284	0.259	0.279	0.285	0.267
		AR-CCE	-	-	0.063	-	-	0.116
		FMTU	0.067	0.247	0.075	0.026	0.294	0.046

1.5 Empirical Application: Urban Geometry in India

In this section I use the proposed inferential method to study the effect of city shape on population density. The data used in this section were originally collected and analyzed in Harari (2020). The shape of a city affects its compactness, where compactness is measured by how convenient its residents travel for daily activities. Ideally, a compact city should look like a circle, whereas cities develop into various shapes for many reasons including geographic constraints. Conceptually, compact cities are attractive to residents because their daily activities operate more efficiently than those in cities that are less compact. This argument suggests the hypothesis that more compact cities should have higher population density.

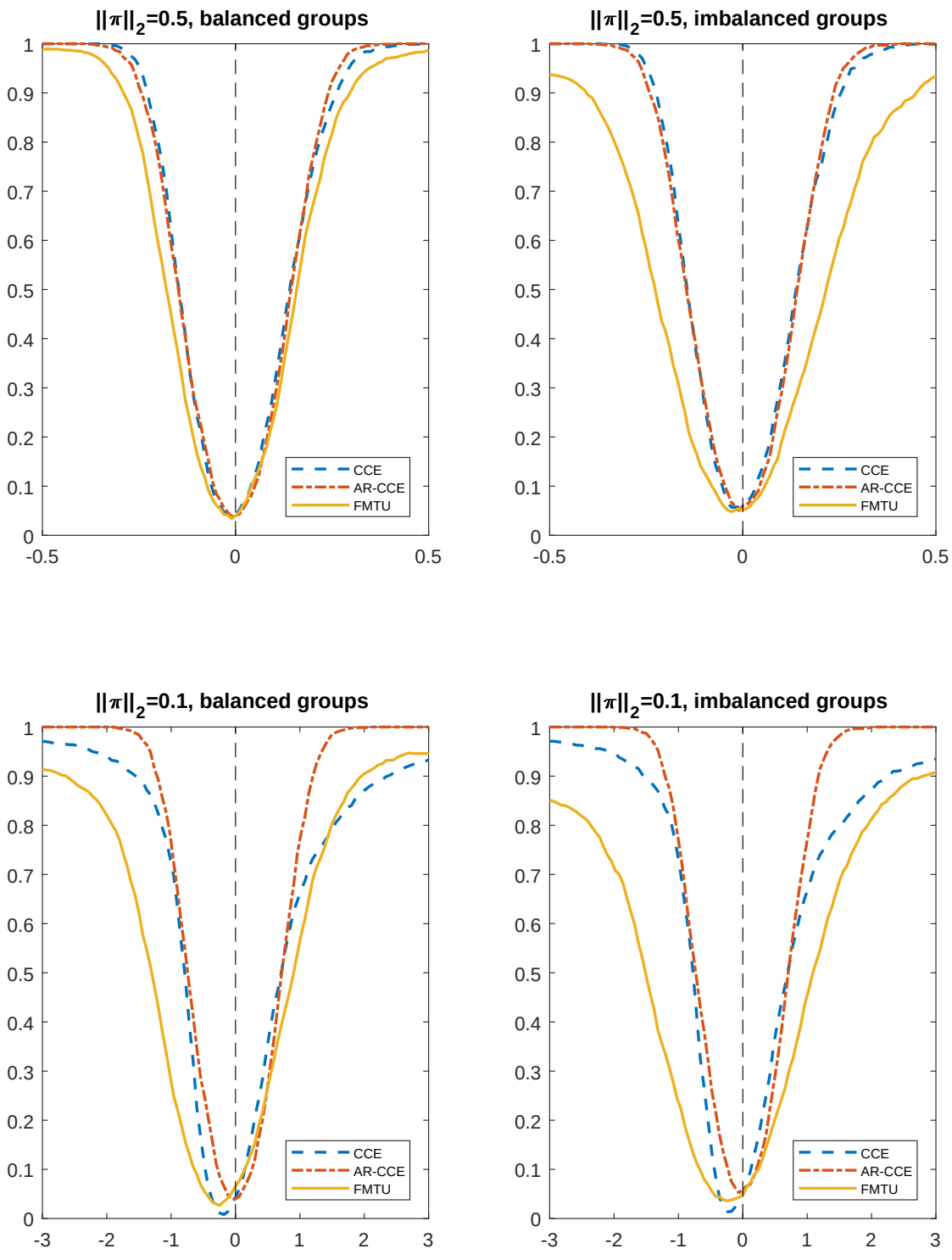


Figure 1.3: Power curves with nominal size $\alpha = 0.05$ and $k = 1$.

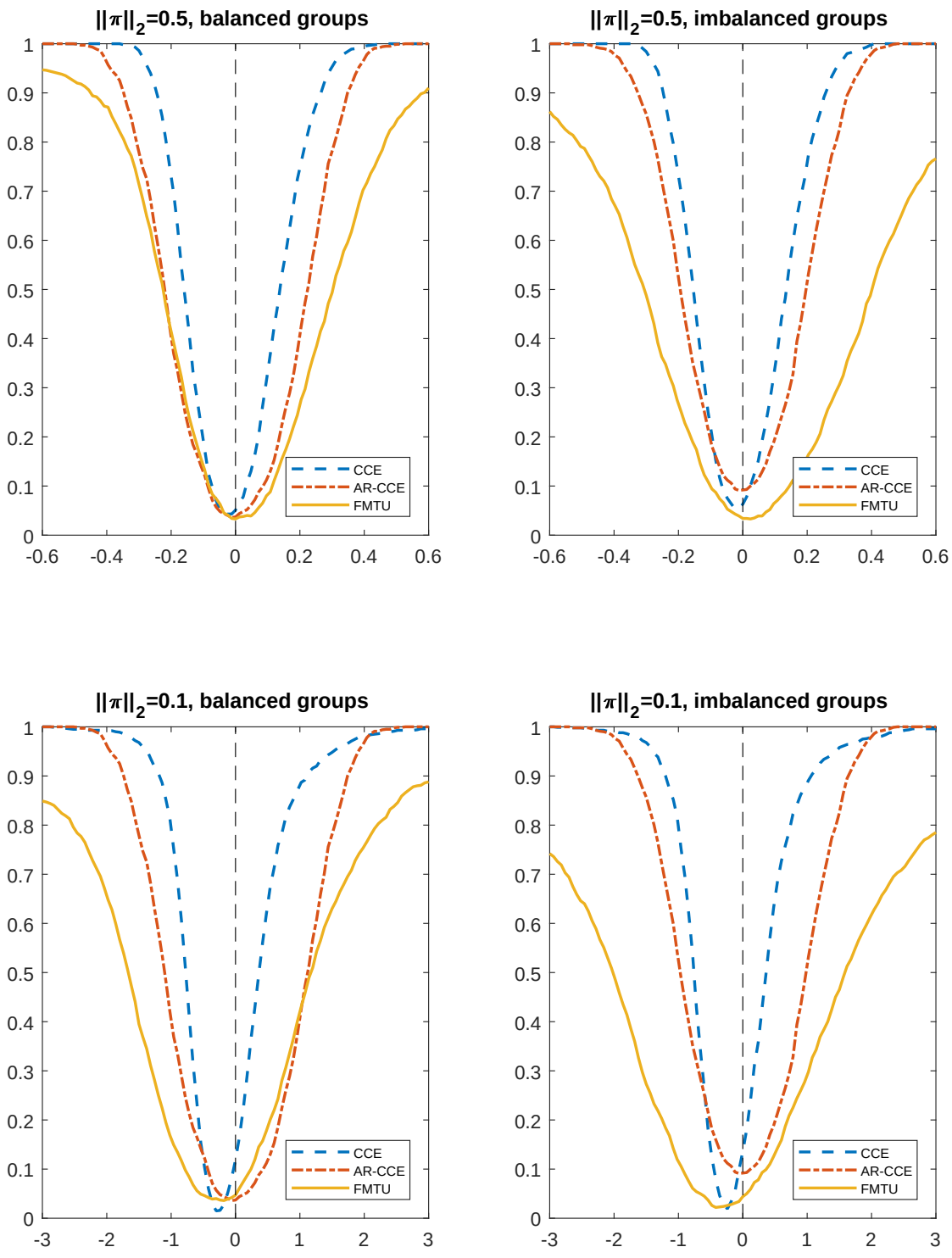


Figure 1.4: Power curves with nominal size $\alpha = 0.05$ and $k = 5$.

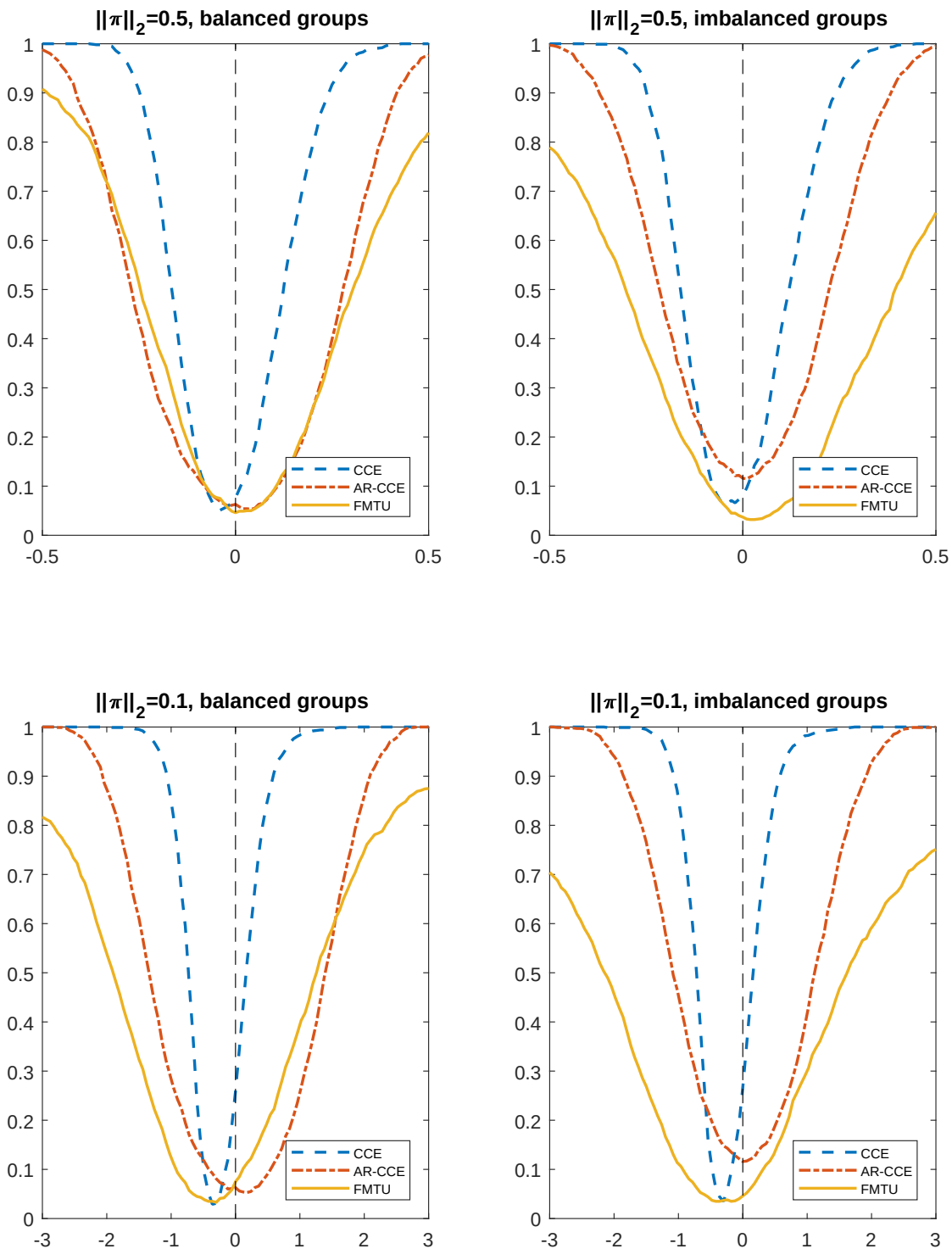


Figure 1.5: Power curves with nominal size $\alpha = 0.05$ and $k = 10$.

However, city shape is highly endogenous because it is the outcome of economic activities. Harari (2020) proposes a solution to this endogeneity problem by utilizing geographic obstacles such as mountains and lakes as an instrument. I apply the method proposed in this chapter, FMTU, in order to obtain a more robust set of empirical results.

1.5.1 Methodology

To facilitate quantitative analysis, Harari (2020) proposes a shape metric that is based on the average distance between any two points in a polygon, in order to measure the compactness of a city. Namely, the Shape index is defined by

$$Shape = \frac{1}{B(B-1)} \sum_{i=1}^{B} \sum_{j \neq i} d_{ij},$$

where i and j stand for two points sampled from interior points of the city, d_{ij} is the Euclidean distance between i and j, and B is the number of sampled points. We consider the *Normalized Shape* obtained by dividing *Shape* by the radius of the Equivalent Area Circle (EAC), where EAC is the circle with the same area as the city. That is, *Normalized Shape* measures how much the city shape is different from a circle.

The instrument is the *Normalized Shape* index for the projected city. Constructing the projected city is a two-step procedure. First, predict the area that a city should occupy in a given year, based on its projected historical population growth. Second, predict the shape of the city given projected area and geographic constraints. We then instrument the shape of the actual city with shape of the potential one.

I consider the same setup as Harari (2020) does. The regression of interest is

$$\Delta Population\ density = \alpha + \beta \Delta Normalized\ shape + U,$$

where the dependent variable is the change in population density from 1951 to 2011, the

endogenous variable is the change in the city shape index from 1950 to 2010, and the instrument is the change in the city-shape index for the projected city expansion from 1950 to 2010. The instrument is the difference of *Normalized Shape* for projected cities. This model can be interpreted as a *difference-in-difference* design with continuous treatment and endogeneity.

I consider the potential dependence among observations by applying the framework suggested by Cao et al. (2019). Namely, I first apply k-medoids to generate a partition of cities using their geographic locations, and then use the given clustering structure to perform the proposed group-based inference method. I use this method to obtain inference results robust to spatial correlation. Factors that affect population density in a city may include climate, culture, economy, personal preferences, etc. Those factors are multidimensional and the natural administrative division[5] does not necessarily capture the underlying dependence structure. That is, cities in neighboring states may be highly correlated in factors that contribute to population density.

The idea of Cao et al. (2019) is to use k-medoids, a clustering algorithm, to generate a partition of observations that helps obtain robust results in group-based inferential methods. Cao et al. (2019) show that the clustering generated by k-medoids satisfies *group-balance* and *diminishing-boundary*. The former, *group-balance*, requires there is no diminishingly small group, and the latter, *diminishing-boundary*, requires across-group dependence is approximately ignorable. The algorithmic details are presented in Algorithm 2.2. I apply k-medoids to generate a clustering of 10 group. The resulting structure is visualized in Figure 1.6. Distances are Euclidean distances based on latitude and longitude coordinates recorded at cities' centroids. Different colors correspond to different clusters in the partition. Marks are plotted at city centroids.

5. India is a federal union comprising 28 states and 8 union territories.

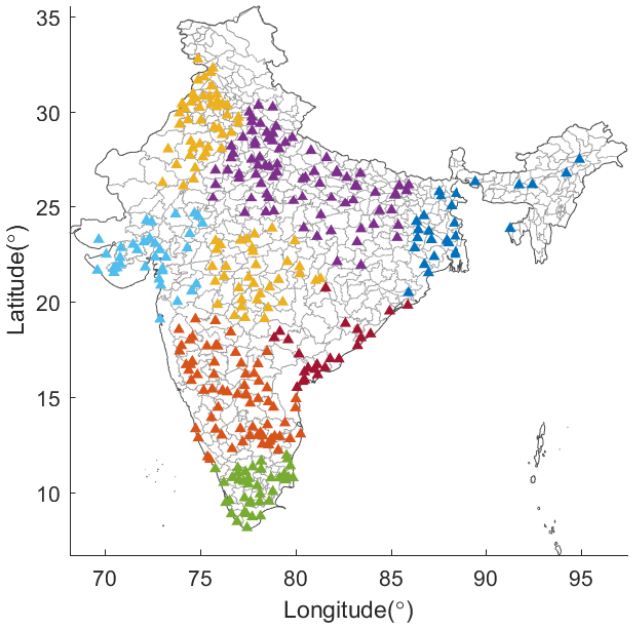


Figure 1.6: Partition of cities in India by k-medoids using 10 clusters.

1.5.2 Results

Table 1.3 compares the original results in Table 8 of Harari (2020) with those obtained from the proposed method. Note that the first-stage t-statistic being 5.311 does not imply we can ignore the instrument strength. Lee et al. (2020) show that in order to have a level-0.05 second-stage test in a single IV model, the first-stage F-statistic needs to exceed 104.7, which translates into a t-statistic of 10.23. Comparing 2SLS and the proposed method of this chapter, the estimates are qualitatively similar (-171.79 vs. -199.26), whereas the standard errors are quite different. Although the new p-value still suggests rejecting the null at some levels such as 0.1, the implication is vastly different from the original p-value, suggesting that including spatial correlation in analysis is crucial.

Table 1.3: Results for the effect of city shape on population density

	Δ Normalized shape	Δ Population density	
	First stage	2SLS	FMUT
	(1)	(2)	(3)
Δ Potential normalized shape	0.0996		
	(0.0188)		
Δ Normalized shape		-171.8	-199.3
		(37.32)	(88.35)
t-stat	5.311	-4.603	-2.255
p-value		0.000	0.051
Observations	351	351	351

Notes: This table reports estimates of the impacts of city shape on population. Column 1 reports the first-stage results. Column 2 reports 2SLS results with the White robust standard error. Column 3 reports results from the truncated unbiased Fama-MacBeth method. The p-value for FMUT is calculated using a Student's t-distribution of 9 degrees of freedom.

1.6 Conclusion

In the setting of IV regression, this chapter proposes an inferential method that is based on the idea of Fama-MacBeth estimation, in order to deal with weak IV and heterogeneous clustering dependence. To overcome the finite-sample bias of IV regression, the group-level estimator is a truncated version of the unbiased IV estimator proposed by Andrews and Armstrong (2017). I give high-level conditions under which the proposed method is asymptotically valid. Asymptotic validity is also shown under both strong and weak IV sequences. Finite-sample performance is shown by simulation. The proposed method is applied to study the effect of city compactness on population density.

CHAPTER 2

INFERENCE FOR DEPENDENT DATA WITH LEARNED CLUSTERS

2.1 Introduction

Conducting accurate econometric or statistical inference with data featuring serial or cross-sectional dependence requires carefully accounting for the respective underlying dependence structure. A variety of methods are available for researchers analyzing dependent data. An important class of inferential methods is the class of *cluster-based* methods. Cluster-based methods work with a partition of observations into clusters. Inference proceeds by treating as negligible any covariance between observations that fall in different clusters, followed by performing appropriate tests for a statistical hypothesis of interest.

Though the term can apply more generally, in empirical economics settings "clustering" is most commonly used in the special case of an $n \times T$ panel data setting, in which individual level observations $i = 1, ..., n$ are clustered together across the T dimension. This setting has been analyzed extensively, and one comprehensive reference is the textbook by Wooldridge (2010). In the case of a panel data setting, there is natural, 'known' way to partition data into clusters, which is justified from an assumption on the data, namely that observations i and j are independent whenever $i \neq j$.

In many settings, however, there is not an immediate 'natural' way to partition the available observations into clusters. This paper considers the case in which a dissimilarity measure is available to the researcher, and this measure is informative about the underlying correlation structure. Specifically, the maintained assumption is that dependence between observable quantities is suitably small when the distance between them is suitably large. The need to cluster general dissimilarity measures can arise in several ways. For instance, this need may arise when observations are equipped with a notion of economic distance.

Examples of applications applying a notion of economic distance include Conley and Dupor (2003), which constructs a notion of economic distance based on a input output tables, or Conley and Topa (2002), who construct a notion of economic distance based on demographic compositions of zip codes. The need for automated partitioning of observations into clusters, may also arise when observations are indexed geographically over an irregular geographical region. To address inferential problems arising in such a way, instead of assuming a natural known partition of observations into clusters, this paper proposes using an unsupervised clustering algorithm to generate a partition.

There are many ways to construct groups from measures of dissimilarity and such problems fall under the category of unsupervised learning; see Hastie et al. (2009) for a general review. Within unsupervised learning, commonly used clustering algorithms include k-medoids, k-means, and hierarchical clustering.

This paper proposes an inferential procedure with three main components for testing a hypothesis H_0. In the first component, given an observed dissimilarity matrix d, a collection of unsupervised clustering methods is used to a collection of partitions $\mathcal{C}$ of the observations into clusters. Next, estimates of size and power for cluster-based procedures over each partition $\mathcal{C}$ in the collection of the previous step are calculated. Finally, a chosen cluster-based inferential procedure is applied using the learned clusters. In the proposed inferential procedure, the total number of clusters in $\widehat{\mathcal{C}}$ may be data-dependent. As a result, the proposed procedure encompasses a method for choosing the level at which to cluster the data.

This paper builds on three commonly used cluster-based methods, which take as input a given known clustering $\mathcal{C}$. Bester et al. (2008) present a simple method for conducting inference using the cluster covariance matrix estimator (CCE) under asymptotics that treat the number of groups as fixed and the number of observations within a group as large. Ibragimov and Müller (2010) provides a formal treatment of the famous Fama-MacBeth

procedure by Fama and MacBeth (1973a), focusing upon properties of t-tests using point estimates from all clusters. Canay et al. (2017) develops a theory of randomization tests under an approximate symmetry assumption, i.e., when the classes of transformations applied to the original data do not change the distribution. Those methods have been shown to be very robust in simulations.

The chosen cluster-based inferential procedure is based a constrained optimization, choosing the procedure which maximizes power while estimated size is nominal. The method is similar to a recent proposal in Mueller and Watson (2014) for selecting a bandwidth in the context of spatial HAC estimation.

The main theoretical results in this paper concern the behavior of cluster-based inference of the procedures defined in Ibragimov and Müller (2010) and Canay et al. (2017) with fixed number of clusters, each of which is determined by an unsupervised data-driven clustering method. The regularity conditions involve moment and mixing rate restrictions, weak homogeneity assumptions on second moments of regressors and unobservables across groups, and restrictions on group boundaries. These moment and mixing conditions are implied by routine assumptions necessary for use of central limit approximations and the required homogeneity is less restrictive than covariance stationarity. Thus the assumptions are no stronger than those routinely made with the plug-in HAC approach or in the Canay et al. (2017) approach.

Intuitively, cluster-based methods are related to truncation or downweighting of autocovariances in traditional time series heteroskedasticity autocorrelation robust inference. Choosing a partition is analogous to choosing a bandwidth. Moreover, the theoretical results in this paper contribute to the growing literature on inference with spatial data; that is, data in which dependence is indexed in more than one dimension. Examples of papers in this literature are Conley (1996), Conley (1999), Kelejian and Prucha (1999), Kelejian and Prucha (2001), Lee (2004) Lee (2007a) Lee (2007b), Jenish and Prucha (2007). Note that

analysis of spatially dependent data is not a trivial extension of results for scalar-indexed (time series). Complications arise due to such concerns as set boundaries being of large order of magnitude relative to set sizes and the number of potential neighbors of any particular point increasing rapidly with the dimension in which dependence increases. This paper provides formal conditions under which inference based on the CCE remains valid in very general settings.

In the course of analysis, this paper develops results about the behavior of the k-medoids clustering algorithm that are important for cluster-based inference and are new to the unsupervised learning literature. More specifically, the partitions produced by k-medoids satisfy suitably defined balance and small boundary conditions. The results do not require a notion of consistency to a true partition. The results about k-medoids complement insights in Bester et al. (2011b). In particular, Bester et al. (2011b) notes that asymptotically correctly sized inferential procedures can be constructed with a small number of clusters under certain regularity conditions, when the clusters have small boundaries and balanced clusters for data indexed in Euclidean spaces with fixed (known) clusters. As a result, this paper shows that unsupervised clustering algorithms may be used to produce clusters which may be used for inference.

This paper also presents simulation evidence on the performance of the proposed procedures in the context of linear panel models. The simulations examine inference in both the context of ordinary least squares (OLS) and instrumental variables (IV) estimation. The simulations illustrate that inference procedures that ignore cross-sectional dependence, such as clustering based on only location, can be severely size distorted. Even modest serial or spatial dependence needs to be accounted for in order to produce reliable inference. The simulation also demonstrates that plug-in spatial-HAC inference procedures may suffer from substantial size distortions. However, when the number of groups is small and correspondingly the number of observations per group is large, the proposed test procedure achieves

nominal size and has good power properties. Finally, the proposed procedure is applied to Condra et al. (2018) to investigate the effect of insurgent attack on voter turnouts.

2.2 Methodology: Inference with Unsupervised Cluster Learning

Consider data given by $\mathscr{D} = \{\zeta_i\}_{i \in \mathsf{X}}$. Here, ζ_i are observable random variables and X is a (spatial) indexing set of cardinality n. This paper assumes that X is equipped with a known dissimilarity measure d, which is an $n \times n$ array of nonnegative real dissimilarities. When added emphasis is helpful, X is written (X, d). The data $\mathscr{D}$ is distributed according to an unknown joint probability distribution $\mathscr{D} \sim \mathrm{P}_0$. The object X will be the main object used to characterize any dependence in the data $\mathscr{D}$ over i. Finally, P_0 is an element of a larger known class of distributions $\mathrm{P}_0 \in \mathbf{P}$.

Let $\mathbf{P}_0 \subseteq \mathbf{P}$ be a subclass of data generating processes of interest. Consider the testing problem

$$H_0 : \mathrm{P}_0 \in \mathbf{P}_0.$$

As a concrete example, this framework includes cases of the problem of testing a coefficient in a linear regression model (i.e. testing $H_0 : \theta_0 = 0$, when θ_0 is a parameter in a linear regression). Then $\mathbf{P}$ consists of all possible data generating processes (DGPs) for the regression data. $\mathbf{P}$ might include DGPs in which regression observations are be correlated with each other. Correspondingly, $\mathbf{P}_0$ consists of all data generating processes considered for which $\theta_0 = 0$. An example of this form is expounded in detail in the next sections.

As discussed in the introduction, failure to account for dependence in $\mathscr{D}$ across $i \in \mathsf{X}$ may lead to substantial size distortion when testing H_0. This paper studies an approach to the above inferential problem with cluster-based inferential procedures, in which the clusters are generated with an unsupervised clustering algorithm.

Let $\mathcal{C} = \{\mathsf{C}_1, ..., \mathsf{C}_G\}$ be a partition of X of cardinality $G \geq 2$. The elements $\mathsf{C}_1, ..., \mathsf{C}_G$

are referred to as clusters. A *Cluster-Based* inferential procedure for testing $H_0 : \mathrm{P}_0 \in \mathbf{P}_0$ is a (possibly random) assignment

$$(\mathscr{D}, \mathcal{C}) \mapsto T$$

where

$$T \in \big\{\text{Fail to Reject, Reject}\big\}.$$

There are many important cluster-based inferential procedures used commonly in econometrics. This paper focuses on the following three cluster-based inferential procedures in the simulation study, empirical example, and formal theoretical development.

IM. The procedure of Ibragimov and Müller (2010).

CRS. The procedure of Canay et al. (2017).

CCE. Inference based on the cluster covariance estimator as described in Bester et al. (2011b).

The procedures are reviewed in the Section 2.6. Each of the above procedures depend on a parameter α, the nominal testing level (in addition to $\mathscr{D}$ and $\mathcal{C}$ as noted above). Each of the above testing procedures has the favorable property of asymptotic nominal size control under respective regularity conditions whenever dependence in observations ζ_i, ζ_j with i, j in different clusters is suitably negligible. The appropriate formal definition of negligible is method-specific. Use the notations

$$T = T_{\mathrm{IM}(\alpha)}, \quad T = T_{\mathrm{CRS}(\alpha)}, \quad \text{and} \quad T = T_{\mathrm{CCE}(\alpha)}$$

to refer to the respective IM, CRS, and CCE testing procedures. Similarly $T_{\bullet(\alpha)}$ is used when the choice of IM, CRS, or CCE is unspecified.

The second important definition is that of an *Unsupervised Clustering Algorithm*, which

is an assignment that returns, to every X, a partition of X given by

$$\mathsf{X} = (\mathsf{X}, d) \mapsto \mathcal{C}.$$

There are also many commonly used unsupervised clustering algorithms. This paper focuses primarily on k-medoids. Other methods included hierarchical clustering, spectral clustering methods, k-means; see Hastie et al. (2009). In particular, all simulations, the empirical example, and theoretical results printed here pertain to k-medoids. A full description of the version of k-medoids used here (included the method of computation) is given in the appendix. The purpose of such clustering algorithms is to partition X into clusters which contain either common features or a small diameter.

By composition, a cluster-based inferential procedure and an unsupervised clustering algorithm are sufficient for defining a statistical test by

$$(\mathscr{D}, \mathsf{X}) \mapsto (\mathscr{D}, \mathcal{C}) \mapsto T.$$

The idea behind using an unsupervised clustering algorithm is that if the dissimilarity d appropriately reflects the dependence in ζ_i, then the resulting partition $\mathcal{C}$ may have the desired property that observations belonging to different clusters exhibit negligible dependence. In the formal analysis that follows, the mixing conditions imply that dependence between ζ_i and ζ_j vanishes as $\mathrm{d}_n(i,j)$ becomes large (in the context of a sequence of X_n). Then, if $\mathcal{C}$ places distant observations (as defined by d) in different clusters, favorable properties of the test T may be anticipated.

The final layer to this paper's proposed testing procedure is that both the cluster-based inferential procedure yielding T and unsupervised clustering procedure yielding $\mathcal{C}$ themselves may depend on the data. This allows for data-dependent choices of tuning parameters which may be inputs into $T, \mathcal{C}$. One important specific consequence of this additional generality in

this paper is that the researcher may explicitly control the final resulting number of clusters G in the partition $\mathcal{C}$, by treating G as a tuning parameter. Let $\mathscr{T}$ be a set of pairs of cluster-based inferential procedures and unsupervised clustering procedures of the form $(T, \mathcal{C})$. Use a pair of circumflexes to denote the possibly data-dependent choice $(\widehat{T}, \widehat{\mathcal{C}}) \in \mathscr{T}$. Then any method for defining $(\widehat{T}, \widehat{\mathcal{C}})$ also defines an assignment

$$(\mathscr{D}, \mathsf{X}) \mapsto \widehat{T}$$

where again

$$\widehat{T} \in \big\{\text{Fail to Reject, Reject}\big\}$$

which in turn fully defines a statistical testing procedure for H_0.

Sensible plans for choosing tuning parameters in statistical testing problems consider simultaneous control of Type-I and Type-II error rates. Let $\text{Err}_{\text{Type-I}}(T, \mathcal{C})$ denote type-I error for $(T, \mathcal{C})$. For a finite set of alternatives $\mathbf{P}_{\text{alt}} \subseteq \mathbf{P}$ disjoint from $\mathbf{P}_0$, let $\text{Err}_{\text{Type-II},\mathbf{P}_{\text{alt}}}(T, \mathcal{C})$ denote average type-II error. Note that in nearly all practical testing settings, the researcher may not know both $\text{Err}_{\text{Type-I}}(T, \mathcal{C})$, $\text{Err}_{\text{Type-II},\mathbf{P}_{\text{alt}}}(T, \mathcal{C})$ exactly. However, in many cases, estimates $\widehat{\text{Err}}_{\text{Type-I}}(T, \mathcal{C}), \widehat{\text{Err}}_{\text{Type-II},\mathbf{P}_{\text{alt}}}(T, \mathcal{C})$ may still be available.

This paper proposes a simple method for choosing data-dependent $(\widehat{T}, \widehat{\mathcal{C}})$ from $\mathscr{T}$. Let α be the nominal testing level for H_0. Consider a setting in which observable estimates $\widehat{\text{Err}}_{\text{Type-I}}(T, \mathcal{C})$ and $\widehat{\text{Err}}_{\text{Type-II},\mathbf{P}_{\text{alt}}}(T, \mathcal{C})$ for some $\mathbf{P}_{\text{alt}}$ exist. Choose $(\widehat{T}, \widehat{\mathcal{C}}) \in \mathscr{T}$ by solving

$$\begin{aligned}(\widehat{T}, \widehat{\mathcal{C}}) &\in \arg\min\ \widehat{\text{Err}}_{\text{Type-II},\mathbf{P}_{\text{alt}}}(T, \mathcal{C}) \\ \text{s.t. } (T, \mathcal{C}) &\in \mathscr{T}\ ,\ \widehat{\text{Err}}_{\text{Type-I}}(T, \mathcal{C}) \leq \alpha.\end{aligned}$$

Such a strategy is particularly well suited for testing with dependent data. In particular, let $G_{\max}$ be a researcher-chosen upper bound on the number of clusters considered. Let $T_{\bullet(a)}$ be as described above for nonnegative a, and let $\mathcal{C}^{(G)}$ be an unsupervised clustering

procedure which yields exactly G clusters. Let

$$\mathscr{T} = \left\{ (T_{\bullet(a)}, \mathcal{C}^{(G)}) : a \in [0, \alpha], G \in \{2, ..., G_{\max}\} \right\}$$

so that $\mathscr{T}$ has a two-dimension parameterization by a and G. Then the optimization defining $\widehat{T}, \widehat{\mathcal{C}}$ is a two-dimensional optimization problem with one constraint (constraining type-I error). The resulting optimization problem is therefore non-degenerate. In this context, corresponding to $\widehat{T}, \widehat{\mathcal{C}}$ are parameters $\widehat{\alpha}$ and $\widehat{G}$. Importantly, $\widehat{G}$ gives researchers guidance on how finely a level to cluster on. The idea of using a constrained optimization problem (rather than maximizing a weighted average of estimated size and power) was previously proposed in Mueller and Watson (2014) in the context of spatial HAC-type estimation.

This proposal departs from several recent methods in the literature which suggest choosing a single tuning parameter (groups in the cluster-based context, bandwidth in the time-series context) to optimize a weighted combination of size distortion and power. See, for instance, Lazarus et al. (2019) and references therein. Instead, this proposal leverages the fact that most commonly used inferential procedures for dependent data depend on two parameters: nominal size and bandwidth.

In the concrete calculations in the following sections, $\widehat{\text{Err}}_{\text{Type-I}}(T, \mathcal{C}), \widehat{\text{Err}}_{\text{Type-II}, \mathbf{P}_{\text{alt}}}(T, \mathcal{C})$ are obtained using a combination of Gaussian QMLE to estimate dependence between relevant observed random variables and simulation to estimate relevant size and power. The simulation studies focus on a simple and intuitive specification with exponential correlation structure which depends on d_n. The theoretical results that follow do not require the consistency of the estimated dependence structure of P_0 in order to control size. As a result, misspecification in the model for dependence leads only to potential loss of power. The implementation details are listed in the appendix.

For convenience of reference in the following sections, the above described procedure is stated under Algorithm 2.1 below.

Algorithm 2.1. *(Inference with Cluster Learning):* $H_0 : \mathrm{P}_0 \in \mathbf{P}_0 \subseteq \mathbf{P}$, level $0 < \alpha < 1$;

Data inputs. $\mathscr{D} = \{\zeta_i\}_{i \in \mathsf{X}}$, (X, d).

Method inputs. $\mathscr{T}$, $(\widehat{T}, \widehat{\mathcal{C}})$.

Output. $\widehat{T} \in \{\text{Fail to Reject, Reject}\}$.

2.3 Empirical Application: The Logic of Insurgent Electoral Violence

In this section, the inferential method proposed in Section 2.2 is applied to study the effect of insurgent attacks on voter turnout using a dataset originally collected and analyzed in Condra et al. (2018). The data consists of district-level observations for 205 voting districts in Afghanistan in 2014. Each observation contains information on direct morning attacks, voter turnout, and district level control variables for 2 separate election rounds.

2.3.1 Empirical specification

The following linear model is specified.

$$Y_{de} = \alpha_0 + \theta_0 Attacks_{de} + W_{de}'\gamma_0 + U_{de},$$

Here, Y_{de} is the turnout in district d and election round e, $Attacks_{de}$ is the number of morning attacks in district d and election round e. The covariates in W_{de} include election round fixed effects, voting hour wind conditions, population, a measure of precipitation and ambient temperature and the average of predawn and morning wind conditions during the pre-election period. U_{de} is an idiosyncratic disturbance term. The parameter θ_0 measures the effect of interest, namely the voter turnout after nearby morning attacks. In order to overcome possible endogeneity, the authors use early morning wind conditions Z_{de} as an instrument for $Attacks_{de}$ in the following first stage regression

$$Attacks_{de} = \mu_0 + \pi_0 Z_{de} + W'_{de}\xi_0 + V_{de}.$$

The analysis in Condra et al. (2018) reports inference for the parameter θ_0 based on cluster standard errors (CCE) at the district level which are robust to within-district correlation (LOC). However, if both the instrumental variable and shocks that affect voting exhibit spatial correlation across districts that are geographically close, then inference based on an assumption of independence might not capture all sources of variability of estimates $\widehat{\theta}$. To test for spatial dependence, a profile of Moran tests was constructed. Weights for all Moran tests are first constructed using positive equal weights for a district's two nearest neighbors (defined by Euclidean distance over latitude and longitude.) Moran tests are run using the estimated observation-specific contributions to the score. Using only observations in the first time period, the calculated Moran I statistic is 4.338 (p-value $< 10_{\text{E}}$-05). Using only observations in the second time period, the calculated Moran I statistic is 2.546 (p-value = 0.011). Using observations in both time periods, treating time periods as very far apart, the calculated Moran I statistic is 5.387 (p-value $< 10_{\text{E}}$-05). An additional Moran test focused on inter-temporal correlation (using only terms for same district in different time periods) delivers a Moran I statistic of 2.755 (p-value = 0.006). The outcomes of all Moran tests are displayed in addition in Table 1. The outcomes of the Moran tests suggest that the spatial and the serial correlation may be important features of this dataset. This section uses the learned cluster-based methodology described in the previous section to perform inference on θ_0 allowing for between-district spatial dependence.

To capture spatial correlation, and implement the procedure defined in the previous section, an unsupervised clustering algorithm is used to partition the districts into clusters. To each pair of distinct districts, d, d', a dissimilarity measure d is defined by

$$\mathrm{d}(d, d') = \|L_d - L_{d'}\|_2 = \sqrt{(\mathrm{lat}_d - \mathrm{lat}_{d'})^2 + (\mathrm{long}_d - \mathrm{long}_{d'})^2}$$

where $L_d = (\text{lat}_d, \text{long}_{d'})$ consists of the latitude and longitude coordinates of the centroid of the district and $\|\cdot\|_2$ denotes Euclidean distance. The data is partitioned using k-medoids with the Euclidean metric over latitude and longitude coordinates of district centroids. Note that because of the nonuniform locations of district centers (i.e. districts are not on a rectangular grid) it is unclear ex ante how to partition them or how many clusters to use. Let $G_{\max} = 8$. This value for $G_{\max}$ is chosen according to $\lceil (NT)^{1/3} \rceil$ and this choice is discussed more in simulation section. For each $G = 2, ..., G_{\max}$, apply k-medoids to the data, yielding a set of partitions $\mathscr{C} = \{\mathcal{C}^{(2)}, ..., \mathcal{C}^{(G_{\max})}\}$. The procedure therefore generates a sequence of group structures $\mathscr{C} = \{\mathcal{C}^{(2)}, \ldots, \mathcal{C}^{(8)}\}$. For any G, the resulting partition is of the form $\mathcal{C}^{(G)} = \{\mathsf{C}_1, \mathsf{C}_2, ..., \mathsf{C}_G\}$

Once the districts have been assigned to clusters, cluster-based inference may proceed. The hypothesis of interest here is $H_0 : \theta_0 = 0$. Let $\mathsf{C} \subseteq \mathsf{X}$ and $\widehat{\theta}_{\mathsf{C}}$ be the least squares estimate (or the IV estimate) of θ_0 using only data corresponding to observations in C. Now define the t-statistic function $t : \mathbb{R}^{\mathcal{C}} \to \mathbb{R}$ such that

$$t(\widehat{\theta}_{\mathcal{C}}) = \frac{G^{-1/2} \sum_{\mathsf{C} \in \mathcal{C}} \widehat{\theta}_{\mathsf{C}}}{\sqrt{(G-1)^{-1} \sum_{\mathsf{C} \in \mathcal{C}} \left(\widehat{\theta}_{\mathsf{C}} - G^{-1} \sum_{\mathsf{C}' \in \mathcal{C}} \widehat{\theta}_{\mathsf{C}'} \right)^2}}$$

for $\widehat{\theta}_{\mathcal{C}} = (\widehat{\theta}_{\mathsf{C}})_{\mathsf{C} \in \mathcal{C}}$. At significance level α, the IM, CRS, and CCE tests relative to partition $\mathcal{C}$ for this particular hypothesis have simple expressions and are given by

$$\begin{aligned}
T_{\text{IM}(\alpha),\mathcal{C}} &= \text{Reject} \quad \text{if} \quad \left| t(\widehat{\theta}_{\mathcal{C}}) \right| > t_{1-\alpha/2, G-1}, \\
T_{\text{CRS}(\alpha),\mathcal{C}} &= \text{Reject} \quad \text{if} \quad \left| t(\widehat{\theta}_{\mathcal{C}}) \right| > \text{quantile}_{1-\alpha/2}(\{|t(h\widehat{\theta}_{\mathcal{C}})|\}_{h \in \mathcal{H}_{\mathcal{C}}}), \\
T_{\text{CCE}(\alpha),\mathcal{C}} &= \text{Reject} \quad \text{if} \quad \left| \frac{\widehat{\theta}_{\mathsf{X}}}{\widehat{V}_{\text{CCE},\mathcal{C}}} \right| > \sqrt{\frac{G}{G-1}} \times t_{1-\alpha/2, G-1},
\end{aligned}$$

where $t_{1-\alpha/2,G-1}$ is the $(1-\alpha/2)$-quantile of t-distribution with $G-1$ degrees of freedom; the set $\{h\widehat{\theta}_{\mathcal{C}}\}_{h \in \mathcal{H}_{\mathcal{C}}}$ is the orbit of the action of $\{\pm 1\}^{\mathcal{C}}$ on $\widehat{\theta}_{\mathcal{C}}$, so that for each h, $h\widehat{\theta}_{\mathcal{C}} \in \mathbb{R}^{\mathcal{C}}$

has Cth component $\pm\widehat{\theta}_{\mathcal{C},\mathsf{C}}$ for some sign in $\{\pm 1\}$; $\widehat{V}_{\text{CCE},\mathcal{C}}$ is the standard cluster covariance matrix estimate.

An estimated 'optimal' estimated number of clusters $\widehat{G}$ for each inferential procedure (CCE, IM, and CRS) is set by selecting G that maximizes simulated average power using critical values that control size within a parametric bootstrap. To calculate power and size for each inferential procedure, a parametric bootstrap in performed over the residuals U_{de}, V_{de} in the above specification. The parametric model for the residuals' covariance is an exponential covariance function that depends on the dissimilarity measure $\mathrm{d}(d, d')$. The following steps describe the procedure to calculate $\widehat{G}$. Full implementation details of each step are provided in the appendix.

Step 1. Let $\widehat{U}_{de}$ and $\widehat{V}_{de}$ be the second stage and first stage estimated residuals from the 2SLS regression of Y_{de} on $Attacks_{de}$ with controls W_{de} and instrument Z_{de}. Estimate by QMLE the parameters of a parametric model for the covariance matrix of $\widehat{U}_{de}$ and the covariance matrix of $\widehat{V}_{de}$ (see Section 2.7).

Step 2. For $* = 1, ..., 1000$, simulate values of U^*_{de} and V^*_{de} from the estimated parametric models in Step 1. Using U^*_{de} and V^*_{de}, simulate values of the dependent variable $Y^*_{de} = \widehat{\alpha} + W'_{de}\widehat{\gamma} + U^*_{de}$ (operating under the assumed null that $\theta_0 = 0$) and values of $Attacks^*_{de}$.

Step 3. For each partition $\mathcal{C}$ that belongs to $\mathcal{C} = \{\mathcal{C}^{(2)}, \ldots, \mathcal{C}^{(8)}\}$, for each $a \in [0, .05]$, compute simulated size $\widehat{\text{Err}}_{\text{Type}-\text{I}}(T_{\text{IM}(a),\mathcal{C}})$, $\widehat{\text{Err}}_{\text{Type}-\text{I}}(T_{\text{CRS}(a),\mathcal{C}})$, $\widehat{\text{Err}}_{\text{Type}-\text{I}}(T_{\text{CCE}(a),\mathcal{C}})$ by conducting the corresponding test $H_0 : \theta_0 = 0$ on each simulated dataset from Step 2. Set $\widehat{a}_{\text{IM},\mathcal{C}}$, $\widehat{a}_{\text{IM},\mathcal{C}}$, $\widehat{a}_{\text{IM},\mathcal{C}}$ to be the largest value $a \in [0, \alpha]$ such that $\widehat{\text{Err}}_{\text{Type}-\text{I}}(T_{\text{IM}(a),\mathcal{C}}) \leq \alpha$, $\widehat{\text{Err}}_{\text{Type}-\text{I}}(T_{\text{CRS}(a),\mathcal{C}}) \leq \alpha$, $\widehat{\text{Err}}_{\text{Type}-\text{I}}(T_{\text{CCE}(a),\mathcal{C}}) \leq \alpha$.

Step 4. For each partition $\mathcal{C}$ that belongs to $\mathcal{C} = \{\mathcal{C}^{(2)}, \ldots, \mathcal{C}^{(8)}\}$, compute simulated average Type-II error, $\widehat{\text{Err}}_{\text{Type-II},\mathbf{P}_{\text{alt}}}(T_{\text{IM}(\widehat{a}_{\text{IM},\mathcal{C}}),\mathcal{C}})$, $\widehat{\text{Err}}_{\text{Type-II},\mathbf{P}_{\text{alt}}}(T_{\text{CRS}(\widehat{a}_{\text{CRS},\mathcal{C}}),\mathcal{C}})$, and $\widehat{\text{Err}}_{\text{Type-II},\mathbf{P}_{\text{alt}}}(T_{\text{CCE}(\widehat{a}_{\text{CCE},\mathcal{C}}),\mathcal{C}})$, using $\widehat{a}_{\text{IM},\mathcal{C}}, \widehat{a}_{\text{CRS},\mathcal{C}}, \widehat{a}_{\text{CCE},\mathcal{C}}$ and by conducting the cor-

responding test $H_0 : \theta_0 = 0$ on each simulated dataset from Step 2. Set $\widehat{G}_{\text{CCE}}$, $\widehat{G}_{\text{IM}}$, $\widehat{G}_{\text{CRS}}$ corresponding to the highest simulated average power. Details on which alternative hypotheses are using in calculating average power are given in Section 2.7. Set

$$\widehat{T}_{IM}(\alpha) = T_{\text{IM}(\widehat{a}_{\text{IM},\mathcal{C}}),\mathcal{C}^{\widehat{G}_{IM}}},$$
$$T_{CRS}(\alpha) = T_{\text{CRS}(\widehat{a}_{\text{CRS},\mathcal{C}}),\mathcal{C}^{\widehat{G}_{CRS}}},$$
$$\widehat{T}_{CCE}(\alpha) = \widehat{T}_{\text{CCE}(\widehat{a}_{\text{CCE},\mathcal{C}}),\mathcal{C}^{\widehat{G}_{CCE}}}.$$

2.3.2 Results

Table 1 reports inferential statistics for the IM, CCE, and CRS cluster-based inferential procedures. Figure 1 displays the resulting partition of the data for $G = 6$. Districts rendered with common markers belong to common clusters $\mathsf{C} \in \mathcal{C}^{(6)}$.

In Table 1, Panel A shows the results of the inferential procedures based on selected clusters. Row labels indicate which procedure is used, including LOC using cluster standard errors at the district level as in Condra et al. (2018), as well as CCE, IM, CRS. For each method, Panel A reports the IV estimate of θ_0, the standard errors, t-statistic, confidence interval using the simulation-adjusted critical values, 95% confidence interval using usual critical values and $\widehat{G}$, which denotes the optimal number of groups selected in each procedure. The optimal number of cluster $\widehat{G}$ is estimated at eight in CCE, which is the maximum allowed by construction, and at six in IM and CRS. Confidence intervals are obtained by inverting relevant hypothesis tests.

Only inference based on LOC using usual critical values rejects the null that $\theta_0 = 0$ at the 5% (C.I. $= [-0.265,\ -0.025]$). This is the inferential procedure considered in Condra et al. (2018). Note that this resulting confidence interval leads to the conclusion that morning

attacks have a negative and significant effect on voter turnout. The LOC confidence interval using the simulation-adjusted critical value is wider (C.I. $= [-0.308,\ 0.018]$) and includes both positive and negative values for θ_0. The IM procedure with adjusted critical values gives confidence interval C.I. $= [-0.896,\ 0.411]$. The CRS procedure with adjusted critical values gives confidence interval C.I. $= [-1.497,\ 0.084]$. Similarly, the CCE procedure with adjusted critical values gives confidence interval C.I. $= [-0.382,\ 0.093]$. In each of the previous three cases, the resulting confidence interval includes the null value $\theta_0 = 0$. Overall, using either simulation-adjusted critical values or usual critical values, inference based on CCE, IM, or CRS does not reject the null that $\theta_0 = 0$.

Table 1, Panel B reports the IV and first-stage estimates of θ_0 for the six subgroups generated using k-medoids as well as for the full sample. Results with 6 groups are reported because 6 is the optimal number of clusters selected by CRS and IM. Note that the column for the full sample reproduces the results in column 4 of table 2 in Condra et al. (2018), which found a significant negative effect of -0.145 (standard error 0.061) of violence over voter turnout.

Table 1 shows that there is less strength in the instrument in the first stage and the IV estimate in the second stage, when looking at the subgroups. For example, in the full sample, the t-statistic associated with the instrument in the first stage is 3.252 whereas in the subsamples the t-statistic is less than 1 in 4 of the six subgroups selected by k-medoids . In the same line, the IV estimate in the second stage is not significant in any of the subgroups selected using k-medoids.

2.4 Simulation

This section conducts a study of the finite sample performance of inference with learned clusters in a series of simulation experiments. The design of the simulation study is largely based on the structure of the empirical example from the previous section. Consider again

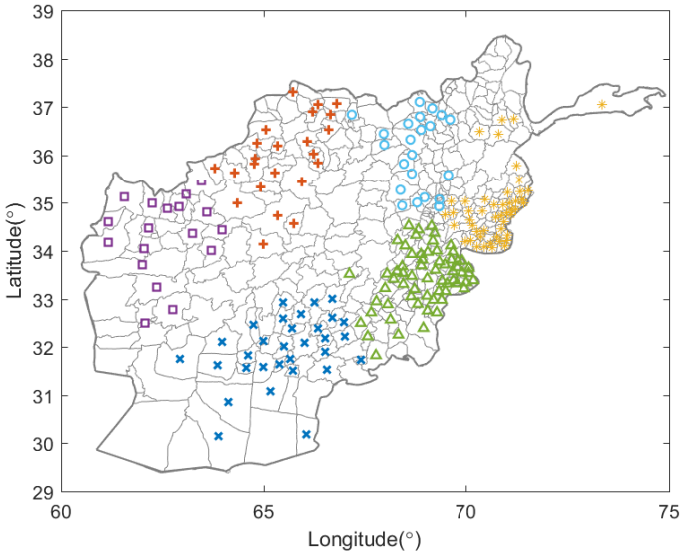


Figure 2.1: Display of partition of districts in Afghanistan by k-medoids using final number of clusters given by $G = 6$. Distances are Euclidean distances based on latitude and longitude coordinates recorded at district centroids. Different marks correspond to different clusters in the partition. Marks are plotted at district centroids.

a $N \times T$ panel from the following process.

$$Y_{de} = \alpha_0 + \theta_0 X_{de} + W_{de}'\gamma_0 + U_{de}.$$

The unknown parameter of interest is θ_0. $W_{de} \in \mathbb{R}^p$ is a vector of control variables with dimension p with effect on the outcome given by the unknown parameter γ_0. Spatial indexing for the indexes d are drawn based on locations of districts in the empirical example in the previous section and are described further below.

Throughout, let $N = 205$, $T = 2$. In addition, $G_{\max}$ is set at $G_{\max} = \lceil (NT)^{1/3} \rceil$ which gives $G_{\max} = 8 = \lceil (205 \times 2)^{1/3} \rceil$. Additional settings with $N = 820$ are provided in the supplemental materials. This model is simulated with several settings which vary the joint distribution of the underlying observable random variables $\{(Y_{de}, X_{de}, W_{de})\}_{d \leq N, e \leq T}$. Specifically, Section 2.4.1 considers OLS estimation of the above panel model under an

orthogonality assumption. Section 2.4.2 assumes an additional observable random variable Z_{de} which is an instrument for X_{de}. The final subsection, Section 2.4.3 studies properties of clustering (for instance, the distribution of the number of chosen group).

The next subsections consider the following inferential procedures.

1. *SK*. Inference based on spatial HAC estimator from Sun and Kim (2015) with optimal bandwidth selection described in Lazarus et al. (2018)

2. *LOC-U*. Inference based on cluster covariance estimator only by location d with the usual critical values (t-distribution with $G-1$ degrees of freedom).

3. *LOC*. Inference based on cluster covariance estimator only by location d with the simulated critical values.

4. *CCE*. Inference based on the cluster covariance estimator described in Section 2.2, i.e., $T = T_{\text{CCE},(\alpha)}$.

5. *IM*. Inference based IM as described in Section 2.2, i.e., $T = T_{\text{IM},(\alpha)}$.

6. *CRS*. Inference based CRS as described in Section 2.2, i.e., $T = T_{\text{CRS},(\alpha)}$.

In all settings, inference is conducted with the null hypothesis $H_0 : \theta_0 = 0$. To obtain the candidate clusterings, k-medoids is applied to the locations $\{L_d\}_{d=1}^N$ over a set of numbers of groups $\{2, \ldots, G_{\max}\}$, which produces $G_{\max} - 1$ partitions. For the CCE method, also consider clustering by location as a benchmark, in which case $G = N$. Implementations details for each of the inferential procedures are provided in the appendix.

2.4.1 OLS simulation

The observable data for the OLS simulation study is given by $\mathscr{D} = \{Y_{de}, X_{de}, W_{de}\}_{d=1,\ldots,N, e=1,\ldots,T}$. The data satisfy $Y_{de} = \alpha_0 + \theta_0 X_{de} + W_{de}'\gamma_0 + U_{de}$. Each observation de is associated with

a spatial location which depends only on $L_d \in \mathbb{R}^2$. Dissimilarity is defined by $\mathrm{d}(de, d'e') = \|L_d - L_{d'}\|_2$. To define a dependence structure over observed random variables, define a function f which depends on positive scalar parameters κ, ρ by

$$f_{\kappa,\rho}((L,t),(L',t')) = \exp\left(-\kappa^{-1}\|L - L'\|_2 - \rho^{-1}|t - t'|\right).$$

The values for $L_d \in \mathbb{R}^2$ used in the simulation study correspond to the locations (district centroid) in the empirical example which has 205 observations. Throughout replications, $\{X_{de}, W_{de}\}_{i=1,\ldots,N,t=1,\ldots,T}$ are held fixed and drawn once and for all from the following distribution.

$$\begin{aligned}
&X_{de} \sim \mathrm{N}(0,1), W_{del} \sim \mathrm{N}(0,1) \\
&\mathrm{corr}(X_{de}, X_{d'e'}) = \mathrm{corr}(W_{del}, W_{d'e'l}) = f_{\kappa,\rho}((L_d, e),(L_{d'}, e')) \\
&\mathrm{corr}(X_{de}, W_{del}) = \mathrm{corr}(W_{del}, W_{dem}) = \nu
\end{aligned}$$

This simulation study consider two settings for the distribution of U_{de}.

A. Homogeneous exponential covariance (*BASELINE*).

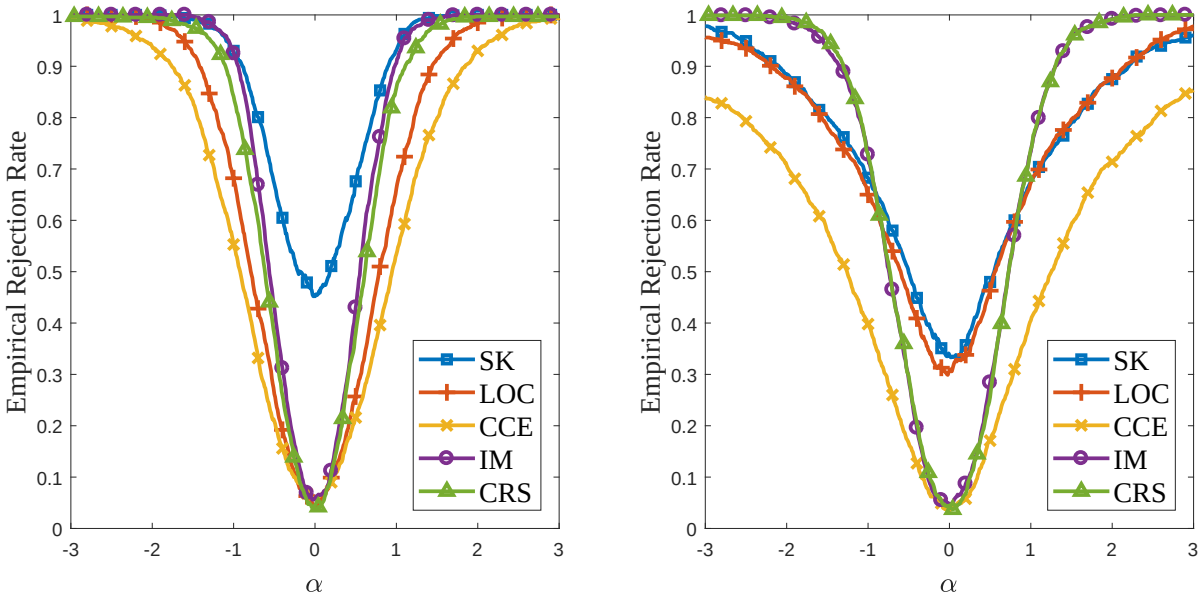
$$\begin{aligned}
&U_{de} \sim \mathrm{N}(0,1) \\
&\mathrm{corr}(U_{de}, U_{d'e'}) = f_{\kappa,\rho}((L_d, e),(L_{d'}, e'))
\end{aligned}$$

B. Spatial auto-regression (*SAR*).

$$\begin{aligned}
&U_{de} = 0.15 \textstyle\sum_{d' \neq d} U_{d'e} \mathbf{1}_{\{\|L_d - L_{d'}\|_2 < 0.3\}} + \varepsilon_{de} \\
&\varepsilon_{de} \sim \mathrm{N}(0,1) \\
&\mathrm{corr}(\varepsilon_{d1}, \varepsilon_{d2}) = \exp(-1)
\end{aligned}$$

Throughout, $\theta_0 = 0$, $\gamma_0 = 0_p = (0, \ldots, 0)'$, $T = 2$, $p = 10$, $\nu = 0.5$, $\kappa = 3$, and $\rho = 1$.

The OLS simulation study calculates the above described inferential procedures on the described data generating process on 1000 simulation replications. The results from this simulation are printed in Table 2.2 and discussed in Section 2.4.4 below. Table 2.2 reports size for each of the 6 inferential procedures considered. Table 2.2 also reports power against alternatives $\theta_0 = -1, -0.5, 0.5, 1$. Table 2.2 also presents statistics about the estimation quality $\widehat{\theta}$, specifically estimation bias and estimation root mean square error. For the SK, LOC-U, LOC, and CCE procedures, $\widehat{\theta}$ is the OLS estimate on the entire sample. For IM and CRS, $\widehat{\theta}$ is given by $\frac{1}{\widehat{G}} \sum_{\mathsf{C} \in \widehat{\mathcal{C}}} \widehat{\theta}_{\mathsf{C}}$. In addition, Figure 2.2 presents a power curve comparing power of the SK, LOC, CCE, IM, and CRS estimators various alternatives for θ_0 ranging between -3 and 3. Table 6 and Figure 4 in the supplemental material report analogous results using a larger sample size of $N = 820$, keeping all other settings the same.



(a) BASELINE (b) SAR

Figure 2.2: OLS power curves

2.4.2 IV simulation

This section conducts a similar simulation study but in an IV setting. The observable data for the IV simulation is given by $\mathscr{D} = \{Y_{de}, X_{de}, W_{de}, Z_{de}\}_{d=1,...,N,e=1,...,T}$. The data satisfy the system of equations

$$Y_{de} = \alpha_0 + \theta_0 X_{de} + W'_{de}\gamma_0 + U_{de},$$
$$X_{de} = \mu_0 + \pi_0 Z_{de} + W'_{de}\xi_0 + V_{de}.$$

In the same way as in the OLS simulation, each observation de is associated with a spatial location which depends only on $L_d \in \mathbb{R}^2$ with dissimilarity is $\mathrm{d}(de, d'e') = \|L_d - L_{d'}\|_2$. A dependence structure over observed random variables is generated using $f_{\kappa,\rho}$, which was defined also in the OLS simulation section. Throughout replications, $\{Z_{de}, W_{de}\}_{i=1,...,N,t=1,...,T}$ are fixed and are drawn once and for all from the following distribution.

$$Z_{de} \sim \mathrm{N}(0,1), W_{del} \sim \mathrm{N}(0,1)$$
$$\mathrm{corr}(Z_{de}, Z_{d'e'}) = \mathrm{corr}(W_{del}, W_{d'e'l}) = f_{\kappa,\rho}((L_d, e), (L_{d'}, e'))$$
$$\mathrm{corr}(Z_{de}, W_{del}) = \mathrm{corr}(W_{del}, W_{dem}) = \nu$$

The following settings for the distribution of the disturbance are considered.

A. Homogeneous exponential covariance (*BASELINE*).

$$U_{de} \sim \mathrm{N}(0,1),\ V_{de} \sim \mathrm{N}(0,1),$$
$$\mathrm{corr}(U_{de}, U_{d'e'}) = \mathrm{corr}(V_{de}, V_{d'e'}) = f_{\kappa,\rho}((L_d, e), (L_{d'}, e'))$$
$$\mathrm{corr}(U_{de}, V_{de}) = 0.8$$

B. Spatial auto-regression (SAR).

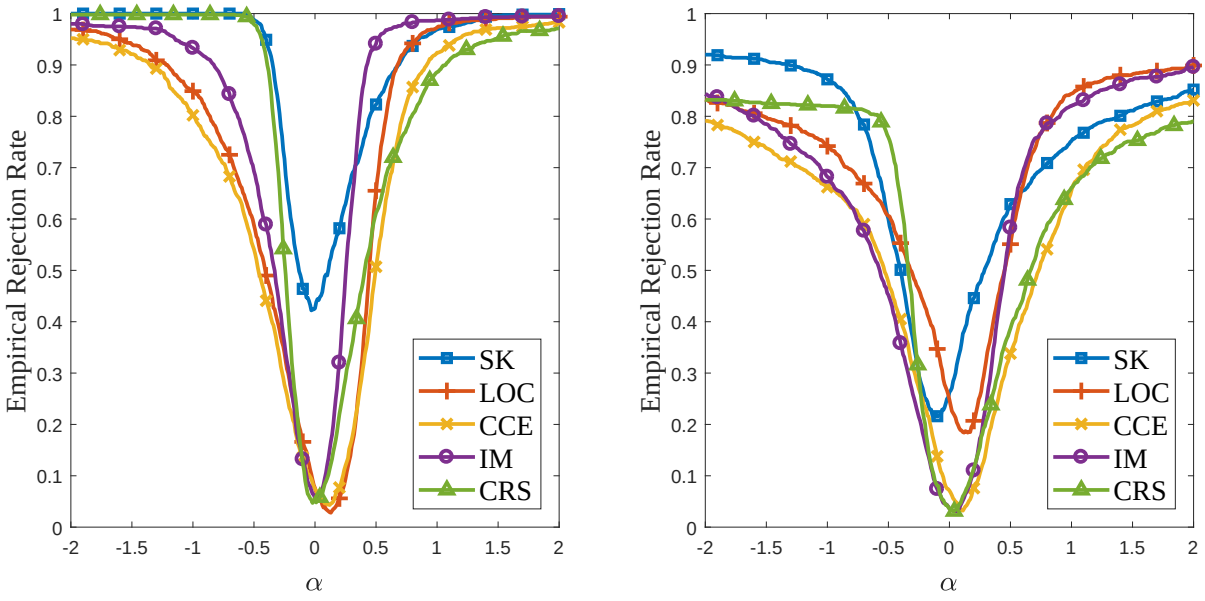
$$
\begin{aligned}
U_{de} &= 0.15 \textstyle\sum_{d' \neq d} U_{d'e} \mathbf{1}_{\{\|L_d - L_{d'}\|_2 < 0.3\}} + \varepsilon_{de} \\
V_{de} &= 0.15 \textstyle\sum_{d' \neq d} V_{d'e} \mathbf{1}_{\{\|L_d - L_{d'}\|_2 < 0.3\}} + \eta_{de} \\
\varepsilon_{de} &\sim \mathrm{N}(0,1),\ \eta_{de} \sim \mathrm{N}(0,1) \\
\mathrm{corr}(\varepsilon_{d1}, \varepsilon_{d2}) &= \mathrm{corr}(\eta_{d1}, \eta_{d2}) = \exp(-1) \\
\mathrm{corr}(\varepsilon_{de}, \eta_{de}) &= 0.8
\end{aligned}
$$

Throughout, $\alpha_0 = \theta_0 = \mu_0 = 0$, $\gamma_0 = \xi_0 = 0_p = (0, \ldots, 0)'$, $\pi_0 = 2$, $T = 2$, $p = 10$, $\nu = 0.5$, $\kappa = 3$, and $\rho = 1$. Note that π_0 controls the relevance of the instrument Z_{de} with $\pi_0 = 0$ indicating an irrelevant instrument.

The IV simulation study calculates the above described inferential procedures on the described data generating process on 1000 simulation replications. The results from this simulation are printed in Table 2.3 and discussed in Section 2.4.4 below. Table 2.3 reports size for each of the 6 inferential procedures considered. Table 2.3 also reports power against alternatives $\theta_0 = -1, -0.5, 0.5, 1$. Table 2.3 also presents statistics about the estimation quality $\widehat{\theta}$, specifically estimation bias and estimation root mean square error. For the SK, LOC-U, LOC, and CCE procedures, $\widehat{\theta}$ is the IV estimate on the entire sample. For IM and CRS, $\widehat{\theta}$ is given by $\frac{1}{\widehat{G}} \sum_{\mathsf{C} \in \widehat{\mathcal{C}}} \widehat{\theta}_{\mathsf{C}}$. In addition, Figure 2.3 presents a power curve comparing power of the SK, LOC, CCE, IM, and CRS estimators various alternatives for θ_0 ranging between -3 and 3. Table 7 and Figure 5 in the supplemental material report analogous results using a larger sample size of $N = 820$, keeping all other settings the same.

2.4.3 Clustering simulation

The final simulation studies more detailed properties of method for choosing the number of clusters $\widehat{G}$ to use. This simulation considers all settings listed above, including both the OLS



(a) BASELINE (b) SAR

Figure 2.3: IV power curves

and IV designs. This simulation study limits consideration of testing procedures to CCE, IM, and CRS. For each $G = 2, ..., 8 = G_{\max}$ in the case of $N = 205$, this simulation records the frequency that $\widehat{G} = G$, as well as size of the procedure using G clusters regardless of whether $\widehat{G} = G$. The results are printed in Table 2.4 and Table 2.5 and discussed in Section 4.4. Tables 8 and 9 in the supplemental material report analogous results using a larger sample size of $N = 820$, keeping all other settings the same.

2.4.4 Discussion of simulation study

This section discusses the results of the above simulation studies.

The OLS simulation study results indicate that in both settings, the SK and LOC-U procedures exhibit substantial size distortion (BASELINE estimated size 0.452 for SK and 0.687 for LOC-U; SAR estimated size 0.338 for SK and 0.638 for LOC-U) . The LOC procedure shows nearly nominal size in BASELINE (estimated size 0.044) but has size distortion in SAR (estimated size 0.307). Note that only in the BASELINE specification, the critical

value of LOC is simulated under the correct specification. The CCE, IM, and CRS procedures achieve sizes that are near nominal (with ranges 0.037–0.059 across all settings). The IM and CRS have similar performance in powers and are more powerful than the CCE in both BASELINE and SAR over all alternatives considered in Table 2.

The IV simulation study results indicate similar findings to the OLS simulation study. In both the BASELINE and SAR settings, the SK and LOC-U procedures exhibit substantial size distortion (BASELINE estimated size is 0.426 for SK and 0.682 for LOC-U; SAR estimated size). The LOC procedure has a near nominal size in BASELINE (estimated size 0.076) but shows size distortion in SAR (estimated size 0.248). The CCE, IM, and CRS procedures achieve sizes that are near nominal (with ranges 0.036–0.075 across all settings). The IM and CRS have similar performance in powers and are more powerful than the CCE in both BASELINE and SAR over all alternatives considered in Table 2. For both BASELINE and SAR, the IM has similar behavior as the CCE near the null hypothesis and higher power away from the null. In general, The CRS procedure has higher power than the CCE on the left ($\theta_0 < -1$) and smaller on the right ($\theta_0 > 2$), partly because of asymmetry in the distributions of respective IV estimators.

In clustering simulation study indicates that across nearly all settings, $G = 8$ has the highest frequencies of being chosen. The exception for this is that the IM method in the case of IV with SAR, where $G = 5$ is chosen with 74.9% chance.[1] The CRS has no power when $G < 6$, so the number of groups is always chosen to be large than or equal to 6.

1. In the supplemental materials, it is shown that the interior solutions for choosing $\widehat{G}$ show up more often when a larger number of locations ($N = 820$) is considered.

2.5 Preliminaries to the Formal Analysis of Inference with Learned Clusters

The analysis of Algorithm 2.1 requires preliminary theoretical groundwork. This section develops regularity conditions on increasing sequences of dissimilarity measures which are sufficient for a mixing central limit theorem to hold. The next section applies the results to analyze Algorithm 2.1 formally for various settings and choices of cluster-based inference procedures.

2.5.1 Conditions on an increasing sequence of dissimilarity measures

The analysis of Algorithm 2.1 relies on asymptotic theory. Consider a sequence of datasets $\mathscr{D}_n$ with $n = 1, 2, 3, ...$ in which the index n in $\mathscr{D}_n$ is identically associated with the sample size. For each n, there is an associated spatial indexing space and dissimilarity $(\mathsf{X}_n, \mathrm{d}_n)$.

The analysis in this paper is based on a condition over dissimilarity matrices. The most basic condition on X_n is that the triangle inequality holds, in other words, each X_n is a genuine metric space. which is related to the notion of Ahlfors regularity in the theory of metric spaces. The asymptotic frame defined by the condition below allows the definition of mixing conditions which are analogous to those in Jenish and Prucha (2009).

Definition 2.1. *Let $|\mathsf{X}|$ be the cardinality of X. A finite metric space X is (C, δ)-finite-Ahlfors regular if $|\mathsf{X}| \vee C^{-1} r^{\delta} \leq |\boldsymbol{B}_{\mathsf{X},r}(i)| \leq C r^{\delta} \wedge 1$ for any $r > 0$, for any x in any X, where $\boldsymbol{B}_{\mathsf{X},r}(x)$ is the r-ball centered at x in the space X.*

(C, δ)-finite-Ahlfors regularity is a modification of the notion of Ahlfors regularity encountered in the theory of metric spaces equipped with a Borel measure (in which case $|\mathbf{B}_{\mathsf{X},r}(i)|$ is replaced by $\mu(\mathbf{B}_{\mathsf{X},r}(i))$ and the conditions $|\mathsf{X}| \vee ...$ and $... \wedge 1$ are dropped). This notion has several advantages, relative to assuming that X_n be a subset of a Euclidean space. First, the definition refers only to intrinsic properties of the space. Second, this simple condition

is sufficient both for realizing mixing central limit theorems, and for analyzing clustering.

Not every dissimilarity measure X_n satisfying (C, δ)-finite-Ahlfors regularity admits an isometric embedding into $\mathbb{R}^\nu$ for some ν. In fact, the condition is not even sufficient to guarantee that X admit a bi-Lipschitz embedding (defined so that the maximum distortion is bounded by a constant) into some $\mathbb{R}^\nu$. However, the condition is strong enough to ensure that X can be "regularized." The new space $(\mathsf{X}, \mathrm{d}^{1-\varepsilon})$, in which the exponent $1-\varepsilon$ is applied element-wise to d, is a metric space for all $\varepsilon \in (0, 1)$ and is called the ε-snowflake of X. The exponent ε serves to regularize the distance d so that it can be embedded into $\mathbb{R}^\nu$ with bounded distortion. This follows from the fact that the doubling constant of finite Ahlfors spaces can be effectively controlled, after which Assoud's embedding Theorem (Assoud (1977)) applies. More details are given in the appendix.

Assumption 2.1. *(Ahlfors Regularity) The sequence of spaces X_n is a uniformly Ahlfors sequence of finite metric spaces.*

Assumption 2.1 defines a spatial asymptotic frame. Note, more explicitly, a sequence X_n of metric spaces is called uniformly Ahlfors if there are nonzero finite constants C, δ which do not depend on n such that each X_n is (C, δ)-finite-Ahlfors regular. The utility of this notion is that gives enough structure to allow analysis of k-medoids techniques analytically as well as derive dependent central limit theorems and laws of large numbers. The condition is also parsimonious and simple to express. Examples of a sequence of metric spaces which satisfy Assumption 2.1 include a sequence $\mathbb{Z}^m \cap \mathsf{Sq}_n$ where Sq_n is the m-dimensional cube of side length $2n$ centered at the origin, or a sequence of annuli $\mathbb{Z}^m \cap \mathsf{A}_n$ where A_n is the m-dimensional annulus of outer radius length $2n$ and inner radius length n centered at the origin. An example of a sequence for which no constituent metric space can be embedded into Euclidean space is the product $\{1, 2, 3, ..., n\} \times \mathsf{W}$ where $(\mathsf{W}, d_{\mathsf{W}})$ is any fixed finite metric space which cannot be embedded into Euclidean space, and the metric on the product is given by the sum of $d_{\mathbb{Z}}$ applied to the first component and d_{W} applied to the second component.

2.5.2 Mixing conditions and a central limit theorem

This section develops a central limit theorem for arrays of random variables on uniformly Ahlfors sequences of metric spaces.

Definition 2.2 (Mixing coefficients)**.** *Let $\mathscr{D}_n = \{\zeta_i\}_{i\in\mathsf{X}_n}$ be an array of random variables on a probability space $(\Omega, \mathscr{F}, \Pr)$ taking values in some measurable space $(V, \mathscr{V})$ with spatial indices given by X_n. Let $\mathscr{A}$ and $\mathscr{B}$ be two (sub-)σ-algebras of $\mathscr{F}$ and let $\alpha(\mathscr{A}, \mathscr{B}) = \sup\{|\Pr(A\cap B) - \Pr(A)\Pr(B)| : A \in \mathscr{A}, B \in \mathscr{B}\}$. For $\mathsf{U}, \mathsf{V} \subseteq \mathsf{X}_n$ let $\alpha_n(\mathsf{U}, \mathsf{V}) = \alpha(\sigma(\zeta_i : i \in \mathsf{U}), \sigma(\zeta_i : i \in \mathsf{V}))$. Let*

$$\alpha_{k,l,n}(r) = \sup\left\{\alpha_n(\mathsf{U},\mathsf{V}), |\mathsf{U}| \le k, |\mathsf{V}| \le l, \mathrm{d}(\mathsf{U},\mathsf{V}) \ge r\right\},$$
$$\bar{\alpha}_{k,l}(r) = \sup_n \alpha_{k,l,n(r)}.$$

The above definition is the same definition as used in Jenish and Prucha (2009). The next condition, Assumption 2.2, relies on the above definition as well as the following fact. Let (X, d) be C, δ-regular. Then $(\mathsf{X}, \mathrm{d}^{3/4})$ has an L-bi-Lipschitz map into $\mathbb{R}^\nu$ where ν and the Lipschitz constant L depend only on C, δ. Define g such that $\nu = g(C, \delta)$.

Assumption 2.2 (Mixing for an Array of Real Scalar Random Variables)**.** *Let $\mathscr{D}_n = \{\zeta_i\}_{i\in\mathsf{X}_n}$ be an array of real scalar random variables on a probability space $(\Omega, \mathscr{F}, \Pr)$ with spatial indices given by X_n satisfying Assumption 2.1 with Ahlfors constants C, δ. Let $\nu = g(\delta, C)$ where $g(\delta, C)$ is the function defined above. For any $\mathsf{C} \subseteq \mathsf{X}$, where $\mathsf{X} = \mathsf{X}_n$ for some n, let $\sigma^2(\mathsf{C}) = \mathrm{var}(\sum_{i\in\mathsf{C}} \zeta_i)$. There exists an array of positive constants $\{\{c_i\}_{i\in\mathsf{X}_n}\}_{n=1}^\infty$ and a positive $\mu > 0$ such that,*

(i) $\mathrm{E}[\zeta_i] = 0$.

(ii) $\lim_{k\to\infty} \sup_{n\ge1} \sup_{i\in\mathsf{X}_n} \mathrm{E}[|\zeta_i/c_i|^{2+\mu}\mathbf{1}_{|\zeta_i/c_i|>k}] = 0$.

(iii) $\sum_{m=1}^\infty \bar{\alpha}_{1,1}(m) m^{\nu\times\frac{\mu+2}{\mu}-1} < \infty$.

(iv) $\sum_{m=1}^{\infty} m^{\nu-1}\bar{\alpha}_{k,l}(m) < \infty$ *for* $k+l \leq 4$.

(v) $\bar{\alpha}_{1,\infty}(m) = O(m^{-\nu-\frac{4}{3}\mu})$.

(vi) $\inf_{n\geq 1} \inf_{\mathsf{C}\subseteq \mathsf{X}_n} |\mathsf{C}|^{-1}(\max_{i\in\mathsf{C}} c_i^{-2})\sigma(\mathsf{C})^2 > 0$.

The conditions are similar to those given for the mixing central limit theorem in Jenish and Prucha (2009) which requires that X_n be a possibly uneven lattice in a finite dimensional Euclidean space with a minimum separation between all points. Note that Assumption 2.1 automatically implies a minimum separation between points. Note also that in the case that X_n embed isometrically into $\mathbb{R}^\nu$ without the need for apply the snowflake regularization construction, then there are only two differences between Assumption 2.2 and the conditions for Corollary 1 of Jenish and Prucha (2009). First, here $\bar{\alpha}_{1,\infty}(m) = O(m^{-\nu-\frac{4}{3}\mu})$ rather than the slightly weaker $\bar{\alpha}_{1,\infty}(m) = O(m^{-\nu-\mu})$. Second, Assumption 2.2(*vi*) now entails an infinum over n and over all subsets $\mathsf{C} \subseteq \mathsf{X}_n$ while Jenish and Prucha (2009) only requires the condition hold for an infinum over n with the particular choice $\mathsf{C} = \mathsf{X}_n$. In terms of notation, here, ζ_i is used rather than the more explicit $\zeta_{i,n}$, because formally the indexes i belong to distinct sets, X_n, for each n. As a result, Assumption 2.2 is a condition on arrays of random variables.

Proposition 2.1. *Suppose that Assumption 2.1 holds for* X_n *and Assumption 2.2 holds for* $\mathscr{D}_n = \{\zeta_i\}_{i\in\mathsf{X}_n}$. *Let* $\mathcal{C}_n$ *be the singleton partition* $\mathcal{C}_n = \{\mathsf{X}_n\}$. *Let* $S_{\mathcal{C}_n}$ *be a random vector with (only) component* $S_{\mathcal{C}_n,\mathsf{X}_n} = \sigma(\mathsf{X}_n)^{-1}\sum_{i\in\mathsf{X}_n}\zeta_i$. *Then there is a probability space* $(\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{\Pr})$ *with random variables* $\tilde{S}_{\mathcal{C}}, \tilde{S}^*_{\mathcal{C}} \in \mathbb{R}^{\mathcal{C}}$ *for every* $\mathcal{C} = \mathcal{C}_n$ *such that* $\tilde{S}_{\mathcal{C}}$ *has the distribution of* $S_{\mathcal{C}}$, $\tilde{S}^*_{\mathcal{C}}$ *has Gaussian distribution with* $\mathrm{var}(\tilde{S}^*_{\mathcal{C},\mathsf{C}}) = \mathrm{var}(\tilde{S}_{\mathcal{C},\mathsf{C}})$ *and* $\tilde{\Pr}$*-almost surely,* $\lim_{n\to\infty} \|\tilde{S}_{\mathcal{C}_n} - \tilde{S}^*_{\mathcal{C}_n}\|_2 = \lim_{n\to\infty} |\tilde{S}_{\mathcal{C}_n,\mathsf{X}_n} - \tilde{S}^*_{\mathcal{C}_n,\mathsf{X}_n}| \to 0$.

The proposition proves a notion of convergence to a Gaussian measure for spatially indexed arrays $\mathscr{D}_n = \{\zeta_i\}_{i\in\mathsf{X}_n}$ which satisfy Assumptions 2.1 and 2.2. It is expressed in

a way so as to match the high-level conditions required in the analysis of IM, CRS, and CCE in the coming sections of this paper. The proof is carried out in detail in the appendix. The proposition can be proven by regularizing X_n using the snowflake construction described above with $\varepsilon = 1/4$, on which the Jenish and Prucha (2009) central limit theorem can then be applied. Thus, the proof first establishes the more familiar-appearing notion of convergence in distribution under the conditions of Proposition 1:

For every $z \in \mathbb{R}$, $\Pr(\sigma(\mathsf{X}_n)^{-1} \sum_{i \in \mathsf{X}_n} \zeta_i \leq z) \to \Phi(z)$ *where* Φ *is the standard Gaussian cumulative distribution function and* $\sigma(\mathsf{X}_n)^2 = \mathrm{var}(\sum_{i \in \mathsf{X}_n} \zeta_i)$.

The proof then applies Theorem 2.19 in van der Vaart (1998). The same argument can be carried out more generally for other $\varepsilon \in (0, 1/2)$, however the single $\varepsilon = 1/4$ snowflake is referenced for added concreteness.

2.5.3 Balance and small common boundary conditions

The key observation in BCH which enables the analysis of the CCE procedure inference is that partitions of X_n with balanced cluster sizes and small boundaries can lead to asymptotically correctly sized inference under mixing conditions. This section formalizes notions of balanced cluster sizes and small boundaries. This section then gives a proposition which shows that under the appropriate regularity conditions, the endpoint of the k-medoids algorithm satisfies these two desired properties.

Assumption 2.3. *The sequence of partitions* $\mathcal{C}_n$ *of* X_n*, for* $n = 1, 2, ...$ *and for which* $|\mathcal{C}_n| = o(n)$*, is asymptotically balanced with small boundaries in the sense of the following conditions.*

(i) The ratios of minimal cluster size to maximal cluster sizes satisfy

$$\liminf_{n\to\infty} \frac{\min_{\mathsf{C}\in\mathcal{C}_n} |\mathsf{C}|}{\max_{\mathsf{D}\in\mathcal{C}_n} |\mathsf{D}|} > 0.$$

(ii) There is a sequence $\bar{r} = \bar{r}(n) \to \infty$ so that

$$\max_{\mathsf{C}\in\mathcal{C}_n} |\{i \in \mathsf{C} : \mathrm{d}(i, \mathsf{X}_n \setminus \mathsf{C}) \leq \bar{r}(n)\}| = o(\min_{\mathsf{C}\in\mathcal{C}_n} |\mathsf{C}|).$$

This definition differs slightly from the definition of small boundary given in BCH. In particular, BCH leverage the fact their spatial domain is a subset of the integer lattice to define neighbor orders for pairs of locations. Their definition of small boundaries entails a bound on the number of first order neighbors from C to $\mathsf{X} \setminus \mathsf{C}$. BCH assume that their given spatial clusters are contiguous and use that fact to bound the number of higher order neighbors from C to $\mathsf{X} \setminus \mathsf{C}$. In this context, there is no available definition of first order neighbor since X_n can be irregular (even non-Euclidean). As a result, this paper works instead with an asymptotic notion of boundary which entails a sequence $\bar{r}$ which allows boundaries to widen as $n \to \infty$. A high-level implication of having asymptotically balanced with small boundaries clusterings is the following proposition.

Proposition 2.2. *Suppose that Assumptions 2.1 and 2.2 hold for $\mathscr{D}_n = \{\zeta_i\}_{i\in\mathsf{X}_n}$ and that Assumption 2.3 holds for partitions $\mathcal{C}_n$ of X_n of bounded size $|\mathcal{C}_n| \leq G_{\max}$ with $G_{\max}$ independent of n. For any $\mathcal{C} = \mathcal{C}_n$ and $\mathsf{C} \in \mathcal{C}$, let $S_{\mathcal{C}} \in \mathbb{R}^{\mathcal{C}}$ have components $S_{\mathcal{C},\mathsf{C}} = \sigma(\mathsf{C})^{-1} \sum_{i\in\mathsf{C}} \zeta_i$. Then there is a probability space $(\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{\Pr})$ with random variables $\tilde{S}_{\mathcal{C}}, \tilde{S}^*_{\mathcal{C}} \in \mathbb{R}^{\mathcal{C}}$ for every $\mathcal{C} = \mathcal{C}_n$ and every n such that $\tilde{S}_{\mathcal{C}}$ has the distribution of $S_{\mathcal{C}}$, $\tilde{S}^*_{\mathcal{C}}$ has Gaussian distribution with independent components with $\mathrm{var}(\tilde{S}^*_{\mathcal{C},\mathsf{C}}) = \mathrm{var}(\tilde{S}_{\mathcal{C},\mathsf{C}})$ and $\tilde{\Pr}$-almost surely, $\lim_{n\to\infty} \|\tilde{S}_{\mathcal{C}_n} - \tilde{S}^*_{\mathcal{C}_n}\|_2 \to 0$.*

The proof is carried out in detail in the supplemental material. Similarly to the proof of

Proposition 1, this proof first establishes that under the conditions of proposition 2,

$$\lim_{n\to\infty} \sup_{\mathsf{C}\neq\mathsf{D}\in\mathcal{C}_n} cov\left(\sigma(\mathsf{C})^{-1}\sum_{i\in\mathsf{C}}\zeta_i,\ \sigma(\mathsf{D})^{-1}\sum_{i\in\mathsf{D}}\zeta_i\right) \to 0.$$

The argument establishing the above fact is related but not identical to arguments in Bester et al. (2011b), (which were previously also given in Jenish and Prucha (2009) and Bolthausen (1982)). Instead of counting points in "shells" around the boundaries of clusters, the proof of Proposition 2 instead relies on the doubling structure implied by the fact that X_n are Ahlfors regular. Both arguments leverage a bound on covariances $\mathrm{cov}(\zeta_i, \zeta_j)$ for sufficiently far spatial as implied by the mixing conditions stated in Assumption 2.2.

The Assumption 2.3 acting as a high-level assumption on clusterings will also be sufficient for asymptotically valid inference in the proposed procedure. In the next section, Assumption 2.3 is verified for k-medoids under an additional convexity assumption. k-medoids is also investigated further in the simulations and in the empirical application below. k-medoids is a popular clustering technique which is related to k-means and in many cases produces very similar clustering results as k-means does. In dealing with a dissimilarity which does not necessarily arise from a Euclidean space, k-means centroids are not necessarily defined. Instead of centroids of clusters as in k-means, k-medoids requires that clusters are defined around elements of X_n called medoids, which must themselves also be elements of X_n. For a detailed description of k-medoids see Hastie et al. (2009). The next proposition derives relevant properties of this version of k-medoids algorithm.

Proposition 2.3. *Assume Assumption 2.2 holds for a sequence X_n. Assume the following additional convexity condition holds. There is a constant K independent of n such that (1) X_n is K-coursely isometric[2] to a subset of a Euclidean space with dimension $N = N(K)$, and (2) for any two point $i, j \in \mathsf{X}_n$ and any $a \in [0,1]$ there is an interpolant $k \in \mathsf{X}_n$ such*

2. $f : (\mathsf{Y}, \mathrm{d}_\mathsf{Y}) \to (\mathsf{Z}, \mathrm{d}_\mathsf{Z})$ is a K-course isometry if $\mathrm{d}_\mathsf{Z}(f(i), f(j)) - K \leq \mathrm{d}_\mathsf{Y}(i,j) \leq \mathrm{d}_\mathsf{Z}(f(i), f(j)) + K$.

that $|\mathrm{d}_n(i,k) - a\mathrm{d}_n(j,k)| \leq K$ *and* $|\mathrm{d}_n(i,k) - (1-a)\mathrm{d}_n(j,k)| \leq K$. *Then the k-medoids algorithm described in the text satisfies Assumption 2.3.*

The proposition does not require that the sequence of resulting partition $\mathcal{C}_n$ of X_n into clusters to converge in any way. Nor does the proposition require there to be any notion of true clusters or a true partition associated to any of the X_n.

2.5.4 Central Limit Theorems in the cases of OLS and IV

Previous propositions show the asymptotic normality and diminishing across-group correlation for *means*. This section gives a similar limiting result in the context of ordinary least squares (OLS) and instrumental variables (IV) estimators. These are the precise central limit theorems which are relevant to the subsequent analysis of cluster-based inferential procedures.

Consider data $\mathscr{D}_n = \{\zeta_i\}_{i \in \mathsf{X}_n}$, with $\zeta_i = (Y_i, X_i, W_i, Z_i)$ in which for each $i \in \mathsf{X}_n$ a linear model holds as follows. $Y_i = \theta_0 X_i + W_i'\gamma_0 + U_i$, where Y_i is a scalar outcome variable, X_i is a scalar variable of interest, U_i is a scalar idiosyncratic disturbance term, W_i is a p-dimensional vector of control variables, and Z_i is a $(p+1)$-dimensional vector of instruments. Consider the IV estimator for θ_0 using only observations $\mathsf{C} \subseteq \mathsf{X}_n$, and, as earlier, denote this by $\widehat{\theta}_{\mathsf{C}}$. Note, the OLS estimate is a special case where $Z_i = X_i$. Let $\mathcal{C}$ be a partition of X_n for some n with G clusters. Define $S_{\mathcal{C}} \in \mathbb{R}^{\mathcal{C}}$ by $S_{\mathcal{C}} \in \mathbb{R}^{\mathcal{C}}$, $S_{\mathcal{C},\mathsf{C}} = \sqrt{|\mathsf{C}|}\,(\widehat{\theta}_{\mathsf{C}} - \theta_0)$ for $\mathsf{C} \in \mathcal{C}$. Note in general that if $\mathcal{C}$ is a partition containing G clusters, then $\mathbb{R}^{\mathcal{C}}$ is a G-dimensional real space with components indexed by $\mathsf{C} \in \mathcal{C}$.

The following proposition holds and verifies the relevant conditions of Theorems 2.1 and 2.2.

Proposition 2.4. *Let* $\mathscr{D}_n = \{(Y_i, X_i, W_i, Z_i)\}_{i \in \mathsf{X}_n}$ *and* $\{(\mathsf{X}_n, \mathrm{d}_n)\}_{n=1}^{\infty}$ *be an array of random elements on a probability space* $(\Omega, \mathscr{F}, \Pr)$ *indexed by* X_n *satisfying* $Y_i = \theta_0 X_i + W_i'\gamma_0 + u_i$. *Suppose Assumption 2.1 holds for the sequence* X_n. *For each* i *in each* X_n, $\mathrm{E}[Z_i U_i]$ *exists*

and is equal to 0 *and* $\mathrm{E}[Z_i X_i']$ *exists and equals some* M_{zx} *which does not depend on* i *or* n *and which has positive eigenvalues. Let* $s_i = Z_i U_i$ *and let* $\eta_i = Z_i X_i' - \mathrm{E}[Z_i X_i']$. *Assumption 2.2 holds for the arrays defined by each component of* s_i *and each component of* η_i. *Suppose that Assumption 2.3 holds for* $\mathcal{C}_n$. *Then there is a probability space* $(\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{\Pr})$ *and random variables* $\tilde{S}_{\mathcal{C}_n}, \tilde{S}^*_{\mathcal{C}_n}$ *such that* $\tilde{S}_{\mathcal{C}_n} =_d S_{\mathcal{C}_n}$, $\tilde{S}^*_{\mathcal{C}_n}$ *is zero-mean and independent Gaussian,* $\liminf_{n\to\infty} \min_{\mathsf{C}\in\mathcal{C}_n} \mathrm{var}(\tilde{S}^*_{\mathcal{C}_n,\mathsf{C}}) > 0$, $\limsup_{n\to\infty} \max_{\mathsf{C}\in\mathcal{C}_n} \mathrm{var}(\tilde{S}^*_{\mathcal{C}_n}) < \infty$ *and* $\|\tilde{S}_{\mathcal{C}_n} - \tilde{S}^*_{\mathcal{C}_n}\|_2 \to 0$ $\tilde{\Pr}$*-almost surely.*

2.6 Analysis of Cluster-Based Inference with Learned Clusters

This section formally develops these applications and states results on the asymptotic validity of the proposed inferential procedures with unsupervised cluster learning. The central limit theorem presented in the earlier section is sufficient (along with additional regularity conditions) for application to show that IM and CRS procedures achieve asymptotically correct size in testing problems associated to OLS and IV estimation problems.

Consider a sequence of spatially indexed datasets $\mathscr{D}_n = \{\zeta_i\}_{i\in\mathsf{X}_n}$. Suppose that each $\mathscr{D}_n$ is distributed according to $\Pr_{0,n}$ which belong to larger classes $\mathbf{P}_n$. Consider the problem of testing a sequence of hypotheses

$$H_{0,n} : P_{0,n} \in \mathbf{P}_{0,n}$$

where $\mathbf{P}_{0,n} \subseteq \mathbf{P}_n$ is a sequence of nulls. This section discusses formal properties of sequences of tests of the from sets $\widehat{T}_n$. Recall from Section 2.2 that $\widehat{T}_n$ is chosen from $(\widehat{T}_n, \widehat{\mathcal{C}}_n) \in \mathscr{T}_n$.

In the cases of IM and CRS inference, for a given partition $\mathcal{C}$, the definition of $T_{\bullet(\alpha)}$ depends on a random vector $S_{\mathcal{C}} \in \mathbb{R}^{\mathcal{C}}$, which depends on $\mathscr{D}_n$ and contain information about departures from $\mathbf{P}_{0,n}$. In the OLS or IV examples discussed in the previous section, $S_{\mathcal{C}}$ with C-th component $S_{\mathcal{C},\mathsf{C}} = \sqrt{|\mathsf{C}|}(\widehat{\theta}_{\mathsf{C}} - 0)$, may be used for testing the hypotheses $H_{0,n} : \theta_0 = 0$.

Other examples of $S_{\mathcal{C}}$ might include rescaled and recentered estimates of linear combinations of coefficients in the OLS or IV model, etc.

The results in the next subsections will imply that, correctly defined, the IM and CRS procedures achieve asymptotically correct size when coupled with an unsupervised clustering algorithm when the data $\mathscr{D}_n, \mathsf{X}_n$ under appropriate regularity conditions. The procedures will be defined and reviewed in the following sections. Proving formal statements about asymptotic size of the procedures follows directly from the results of IM and CRS if each $\mathscr{C}_n$ contains only one partition $\mathcal{C}_n$ and $|\mathcal{C}_n| = G$ for some common integer G independent of n. To show that the conclusions of Theorems 2.1 and 2.2 continue to hold when $T_{\bullet(\alpha)}$ is replaced by $\widehat{T}_{\bullet}$ in the cases of IM and CRS, it is required to show that the action of selecting a partition has negligible effect on T.

This section derives properties of $\widehat{T}_n$ when the sequence of sets $\mathscr{T}_n$ is sufficiently regular. Specifically, results here are stated when $\mathscr{T}_n$ consists of tests of the form $T_{\mathrm{IM}(\alpha)}$ and $T_{\mathrm{CRS}(\alpha)}$ and partitions $\mathcal{C}_n$ satisfying Assumption 2.2. Discussion of $T_{\mathrm{CCE}(\alpha)}$ is also provided. For any n let $\mathscr{C}_n$ be its collection of partitions $\mathscr{C}_n = \{\mathcal{C} : (T, \mathcal{C}) \in \mathscr{T}_n \text{ for some } T\}$. The next assumption is a high-level condition which almost immediately guarantees that the conclusions of the previous two theorems continue to hold under a selected $\widehat{\mathcal{C}}$. Discussion on more primitive assumption is given afterwards.

Assumption 2.4. *For each n, there is a set $\mathscr{C}_{0,n} \subseteq \mathscr{C}_n$ such that as $n \to \infty$,*

(i) $\Pr(\widehat{\mathcal{C}}_n \notin \mathscr{C}_{0,n}) = o(1)$,

(ii) $\sup_{\mathcal{C} \in \mathscr{C}_{0,n}} |\Pr(\widehat{T}_{\bullet(\alpha),\mathcal{C}} = \text{Reject} | \widehat{\mathcal{C}}_n = \mathcal{C}) - \Pr(T_{\bullet(\alpha),\mathcal{C}} = \text{Reject})| = o(1)$.

Note that if $\mathscr{C}_{0,n}$ exist, then each $\mathscr{C}_{0,n}$ may be taken finite and each $\mathcal{C} \in \mathscr{C}_{0,n}$ may be taken to have $\Pr(\widehat{\mathcal{C}}_n = \mathcal{C}) > 0$. The method for selecting $\widehat{\mathcal{C}}$ in this paper is through a combination of QMLE and constrained minimization of a loss function (defined as estimated weighted average power) as described in Section 2.

Under sufficient regularity conditions, the QMLE procedure defined in the appendix is consistent for a fixed object which may depend on n. If the loss function is sufficiently regular, than $\mathscr{C}_{0,n}$ is a singleton. Alternatively, if a single fixed G is used in conjunction with k-medoids clustering, then Assumption 2.4 is automatically satisfied.

2.6.1 Definition and analysis for IM with learned clusters

The IM procedure is given as follows. Consider any $S \in \mathbb{R}^{\mathcal{C}}$ and let G be the number of clusters in $\mathcal{C}$. Define the two quantities

$$\bar{S} = G^{-1} \sum_{\mathsf{C} \in \mathcal{C}} S_{\mathsf{C}}, \ \mathrm{sd}(S) = \sqrt{(G-1)^{-1} \sum_{\mathsf{C} \in \mathcal{C}} (S_{\mathsf{C}} - \bar{S})^2}$$

In addition, define a t-statistic by

$$t(S) = \frac{\sqrt{G}\bar{S}}{\mathrm{sd}(S)}.$$

At significance level α, the IM test relative to partition $\mathcal{C}$ is given by

$$T_{\mathrm{IM}(\alpha),\mathcal{C}} = \text{Reject} \quad \text{if} \quad |t(S_{\mathcal{C}})| > t_{1-\alpha/2,G-1},$$

where $t_{1-\alpha/2,G-1}$ is the $(1-\alpha/2)$-quantile of t-distribution with $G-1$ degrees of freedom. Then define for each n,

$$\widehat{T}_{\mathrm{IM}(\alpha),n} = T_{\mathrm{IM}(\widehat{\alpha}),\widehat{\mathcal{C}}}$$

where $(T_{\mathrm{IM}(\widehat{\alpha}),\widehat{\mathcal{C}}}, \widehat{\mathcal{C}}) \in \mathscr{T}_n$.

Assumption 2.5 (Regularity Conditions for IM)**.**

(i) The significance level satisfies $\alpha \leq 2\Phi(-\sqrt{3}) = 0.083...$, Φ *the standard Gaussian distribution function.*

(ii) $\liminf_{n\to\infty} \inf_{\mathsf{C}\in\mathcal{C}\in\mathscr{C}_n} \mathrm{var}(S_{\mathcal{C}}) > 0$ *and* $\limsup_{n\to\infty} \sup_{\mathsf{C}\in\mathcal{C}\in\mathscr{C}_n} \mathrm{var}(S_{\mathcal{C},\mathsf{C}}) < \infty$.

(iii) There is a probability space $(\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{\Pr})$*, random variables* $\tilde{S}_{\mathcal{C}}, \tilde{S}^*_{\mathcal{C}} \in \mathbb{R}^{\mathcal{C}}$ *for every* $\mathcal{C} \in \mathscr{C}_n$ *and every* n*, and* $\tilde{U} \in [0,1]$ *all defined on* $\tilde{\Pr}$*.* $\tilde{S}_{\mathcal{C}}$ *has the distribution of* $S_{\mathcal{C}}$*,* $\tilde{S}^*_{\mathcal{C}}$ *has Gaussian distribution with independent components with* $\operatorname{var}(S_{\mathcal{C},\mathsf{C}}) = \operatorname{var}(\tilde{S}^*_{\mathcal{C},\mathsf{C}})$ *and* $\tilde{U}$ *is uniformly distributed. Furthermore,* $\lim_{n\to\infty} \sup_{\mathcal{C}\in\mathscr{C}_n} \|\tilde{S}_{\mathcal{C}} - \tilde{S}^*_{\mathcal{C}}\|_2 \to 0$ *almost surely.*

Assumption 2.5(i) is also needed in Ibragimov and Müller (2010). Assumptions 2.5(ii) and (iii) are standard non-degeneracy and regularity conditions for $S_{\mathcal{C}}$. Assumption 2.5(iii) is simultaneously a high level central limit theorem and almost sure representation theorem. Other conditions are possible. The analysis of the CRS procedure in the next section requires an almost sure representation result, and thus it is assumed here to maintain parallel exposition. Assumption 2.5 was verified for the OLS and IV models in the previous section with partitions generated by k-medoids with fixed G.

Theorem 2.1. *Suppose that Assumptions 2.1 and 2.2 hold for* $\mathscr{D}_n = \{\zeta_i\}_{i\in\mathsf{X}_n}$ *and that Assumption 2.3 holds uniformly (with the same set of* O *constants and* o *sequences) for partitions* $\mathcal{C}_n \in \mathscr{C}_n$ *on* X_n*. Suppose Assumption 2.4 holds for* $\widehat{T}_{\mathrm{IM}(\alpha),n}$ *and for some fixed* $\alpha > 0$*. Suppose Assumption 2.5 holds for* $S_{\mathcal{C}} \in \mathbb{R}^{\mathcal{C}}$*. Then under* $H_{0,n}$*,*

$$\limsup_{n\to\infty} \Pr(\widehat{T}_{\mathrm{IM}(\alpha),n} = \text{Reject}) \leq \alpha.$$

2.6.2 Definition and analysis for CRS with learned clusters

Another possible statistical test for $H_{0,n}$ is that given by $T_{\mathrm{CRS}(\alpha),n}$. This test was developed in Canay et al. (2017) and is a randomization test which requires that the distribution of the observed data exhibits approximate symmetry. The procedure defining $T_{\mathrm{CRS}(\alpha),n}$ depends on the availability of a real-valued test statistic W, which is defined such that large values of W provide evidence against H_0. Given a partition $\mathcal{C}$ of X_n, the CRS procedure also depends on the availability of a finite group $\mathcal{H}_{\mathcal{C}}$ symmetries of the data which act within component C of $\mathcal{C}$. These symmetries are formalized by a group action $\mathscr{D}_n \mapsto h\mathscr{D}_n$ for all $h \in \mathcal{H}_{\mathcal{C}}$.

In the context of cluster-based inference in this paper, the only type of group action of interest is a set of signs $\mathcal{H}_{\mathcal{C}} = \{-1, +1\}^{\mathcal{C}}$ that operates within clusters. For concreteness, the group elements of $\mathcal{H}_{\mathcal{C}}$ are simply referred to as signs. Elements, h, which are tuples of signs $h_{\mathsf{C}} \in \{-1, +1\}$ indexed by $\mathsf{C} \in \mathcal{C}$ will act by ordinary multiplication, with the component h_{C} multiplying variables indexed by observations $i \in \mathsf{C} \subseteq \mathsf{X}_n$. The action on the data is defined in such a way that

$$\mathscr{D}_n \mapsto h\mathscr{D}_n$$

induces a map

$$S_{\mathcal{C}} \mapsto hS_{\mathcal{C}}$$

which respects the original action in the sense $S_{\mathcal{C}}(h\mathscr{D}_n) = hS_{\mathcal{C}}(\mathscr{D}_n)$. Within the cluster-based inference framework, the induced maps have the additional structure that the action of h on $S_{\mathcal{C}}$ decomposes componentwise so that $S_{\mathcal{C},\mathsf{C}} \mapsto h_{\mathsf{C}} S_{\mathcal{C},\mathsf{C}}$.

In the context of the OLS and IV models, one action $\mathscr{D}_n \mapsto h\mathscr{D}_n$ is given by

$$(Y_i, X_i, W_i, Z_i) \mapsto (Y_i, h_{\mathsf{C}} X_i, W_i, Z_i) \quad \text{for } i \in \mathsf{C} \in \mathcal{C}_n$$

where $h_{\mathsf{C}} X_i$ is ordinary multiplication of $h_{\mathsf{C}} \in \{-1, +1\}$ with X_i. Note that for $S_{\mathcal{C}}$ as defined previously, such action does imply that $S_{\mathcal{C},\mathsf{C}} \mapsto h_{\mathsf{C}} S_{\mathcal{C},\mathsf{C}}$. I.e., multiplying all variables y_i by a sign results in multiplying the $\sqrt{|\mathsf{C}|}(\widehat{\theta}_{\mathsf{C}} - \theta_0)$ by the same sign under $H_0 : \theta_0 = 0$.

Let w be a statistic depends on the data $\mathscr{D}_n$ only through $\mathcal{S}_{\mathcal{C}}$. Therefore, with slight abuse of notation, write $w(\mathscr{D}_n) = w(S_{\mathcal{C}})$. In all implementations in this paper, w is chosen to be the absolute value of the mean $w(S) = |\bar{S}|$. Let $M = |\mathcal{H}_{\mathcal{C}}|$ and let $w^1(S_{\mathcal{C}}) \leq w^2(S_{\mathcal{C}}) \leq \cdots \leq w^M(S_{\mathcal{C}})$ denote the order statistics of the orbit $\{w(h\mathscr{D}_n) : h \in \mathcal{H}\}$. Let $k = M(1-\alpha)$, $M^+(X) = |\left\{1 \leq j \leq M : w^j(S_{\mathcal{C}}) > w^k S_{\mathcal{C}}\right\}|$ and $M^0(X) = |\left\{1 \leq j \leq M : w^j(X) = w^k(S_{\mathcal{C}})\right\}|$.

Let $\tilde{a}\left(S_{\mathcal{C}}\right)=\frac{M\alpha-M^{+}(S_{\mathcal{C}})}{M^{0}(S_{\mathcal{C}})}$. A randomization test is given by:

$$T_{\mathrm{CRS}(\alpha),n}=\text{Reject}\quad\text{if}\quad w(S_{\mathcal{C}})>w^{k}(S_{\mathcal{C}})\quad\text{or}\quad\Big(w(S_{\mathcal{C}})=w^{k}(S_{\mathcal{C}})\text{ and }\tilde{a}(S_{\mathcal{C}})=1\Big).$$

In Canay et al. (2017) it was shown that if (i) $S_{\mathcal{C}}\xrightarrow{d}S^{*}$ for some fixed S^{*}, (ii) $hS_{\infty}=_{d}S^{*}$ for all h, (iii) for distinct h,h', either $w\circ h=w\circ h'$ or $\Pr(w(hS^{*})\neq w(h'S^{*}))=1$, (iv) w is continuous and the action of h is continuous for each h, then $\Pr(\widehat{T}_{\mathrm{CRS}(\alpha),n}=\text{Reject})\to\alpha$.

Assumption 2.6 (Regularity Conditions for CRS)**.**

(i) $\liminf_{n\to\infty}\inf_{\mathsf{C}\in\mathcal{C}\in\mathscr{C}_n}\mathrm{var}(S_{\mathcal{C},\mathsf{C}})>0$ *and* $\limsup_{n\to\infty}\sup_{\mathsf{C}\in\mathcal{C}\in\mathscr{C}_n}\mathrm{var}(S_{\mathcal{C},\mathsf{C}})<\infty$.

(ii) There is a probability space $(\tilde{\Omega},\tilde{\mathscr{F}},\tilde{\Pr})$ *with random variables* $\tilde{S}_{\mathcal{C}},\tilde{S}^{*}_{\mathcal{C}}\in\mathbb{R}^{\mathcal{C}}$ *for every* $\mathcal{C}\in\mathscr{C}_n$ *and every* n*, and* $\tilde{U}\in[0,1]$ *all defined on* $\tilde{\Omega}$*.* $\tilde{S}_{\mathcal{C}}$ *has the distribution of* $S_{\mathcal{C}}$*,* $\tilde{S}^{*}_{\mathcal{C}}$ *has Gaussian distribution with independent components with* $\mathrm{var}(S_{\mathcal{C},\mathsf{C}})=\mathrm{var}(\tilde{S}^{*}_{\mathcal{C},\mathsf{C}})$ *and* $\tilde{U}$ *is uniformly distributed independent of* $\tilde{S}^{*}_{\mathcal{C}}$ *and* $\tilde{S}_{\mathcal{C}}$*. Furthermore,* $\lim_{n\to\infty}\sup_{\mathcal{C}\in\mathscr{C}_n}\|\tilde{S}_{\mathcal{C}}-\tilde{S}^{*}_{\mathcal{C}}\|_2\to 0$ $\tilde{\Pr}$*-a.s..*

(iii) $h\tilde{S}^{*}_{\mathcal{C}}$ *has the same distribution as* $\tilde{S}^{*}_{\mathcal{C}}$ *for all* $h\in\mathcal{H}_{\mathcal{C}}$*, for all* $\mathcal{C}$ *in all* $\mathscr{C}_n$*; for distinct* h,h'*, either* $w\circ h=w\circ h'$ *or* $\tilde{\Pr}(w(h\tilde{S}^{*}_{\mathcal{C}})\neq w(h'\tilde{S}^{*}_{\mathcal{C}}))=1$*;* w *is continuous and the action of* h *is continuous for each* h *in each* $\mathcal{H}_{\mathcal{C}}$*. In addition, either*

a. $|\mathscr{C}_n|$ *and* $\max_{\mathcal{C}\in\mathscr{C}_n}|\mathcal{C}|$ *are each bounded by a constant independent of* n*; or*

b. For any sequence $\check{G}_n\to\infty$ *sufficiently slowly, the following hold. There is a sequence* $\check{\delta}_n\to 0$ *and sets* $\mathcal{A}_{\mathcal{C}}\subseteq\mathbb{R}^{\mathcal{C}}$ *for each* $\mathcal{C}\in\mathscr{C}_n$ *with* $|\mathcal{C}|>\check{G}_n$ *which are closed under the action of* $\mathcal{H}_{\mathcal{C}}$ *such that,* $\lim_{n\to\infty}\inf_{\mathcal{C}\in\mathscr{C}_n,|\mathcal{C}|>\check{G}}\tilde{\Pr}(\mathcal{A}_{\mathcal{C}})=1$*,* $w(\cdot)$ *can be renormalized to have Lipschitz constant 1 respect to the Euclidean norm on all* $\mathcal{A}_{\mathcal{C}}$*, and it holds that* $\sup_{\mathcal{C}\in\mathscr{C}_n}\sup_{r\in\mathbb{R}}\Pr_{\mathcal{C}}(|r-w(h\tilde{S}^{*}_{\mathcal{C}})|<\check{\delta}_n)\to 0$ *where* $\Pr_{\mathcal{C}}$ *be the uniform probability measure over* $h\in\mathcal{H}_{\mathcal{C}}$*. Finally,* $|\{\mathcal{C}\in\mathscr{C}_n:|\mathcal{C}|\leq\check{G}_n\}|$ *depends only* $\check{G}_n$*.*

Assumption 2.6(ii) is a high-level uniform central limit theorem, which is similar to what was required for the analysis of IM, with the only difference being that $\tilde{U}$ is also defined on $\tilde{\Omega}$. Assumption 2.6(iii) is a high-level assumption about anticoncentration of $w \circ h$ as h ranges over $\mathcal{H}_{\mathcal{C}}$ after w has been suitably renormalized. The next proposition gives an example in which Assumption 2.6(iii) holds.

Proposition 2.5. *For $S \in \mathbb{R}^{\mathcal{C}}$, let $w(S) = t(S)$. Let $S^*_{\mathcal{C}}$ be a Gaussian random variables with independent components with variances bounded from below and above (as in Assumption 2.6(i)). Suppose that $\mathscr{C}_n = \{\mathcal{C}^{n,(2)}, ..., \mathcal{C}^{n,(G_{\max})}\}$ where $G_{\max}$ may depend arbitrarily on n. Then Assumption 2.6(iii)*a *is satisfied if $G_{\max}$ is bounded and Assumption 2.6(iii)*b *is satisfied if $G_{\max} \to \infty$.*

The above proposition verifies that the test statistic, w, defined in the text above and used for the empirical example and simulation study satisfy the anticoncentration requirement of Assumption 2.6. The next theorem shows that under the stated regularity conditions, the procedure proposed in this paper in conjunction with CRS cluster-based inference achieves asymptotically correct size.

Theorem 2.2. *Suppose that Assumptions 2.1 and 2.2 hold for $\mathscr{D}_n = \{\{\zeta_i\}_{i \in \mathsf{X}_n}, (\mathsf{X}_n, d_n)\}_{n=1}^{\infty}$ and that Assumption 2.3 holds uniformly (with the same set of O constants and o sequences) for partitions $\mathcal{C}_n \in \mathscr{C}_n$ on X_n. Suppose Assumption 2.4 holds for $\widehat{T}_{\mathrm{CRS}(\alpha),n}$ and for some fixed $\alpha > 0$ and Assumption 2.6 holds for $S_{\mathcal{C}} \in \mathbb{R}^{\mathcal{C}}$. Then under $H_{0,n}$,*

$$\lim_{n \to \infty} \sup \Pr(\widehat{T}_{\mathrm{CRS}(\alpha),n} = \text{Reject}) \leq \alpha.$$

The theorem is proven in the appendix. In the case of bounded partitions $\mathcal{C}$ (ie. $G_{\max} \geq \sup_n \sup_{\mathcal{C} \in \mathscr{C}_n} |\mathcal{C}|$ for some $G_{\max}$ independent of n), the proof of Theorem 2.2 follows CRS. The argument for handling sequences of partitions of increasing size is new relative to CRS and depends strongly on the anticoncentration condition, Assumption 2.6(iii).

2.6.3 Discussion for BCH with learned clusters

An alternative to the IM and CRS procedures above is the CCE estimator as described in BCH. Note that the CCE estimator is the standard cluster standard error estimator. In the context of the linear model in previous sections, the CCE test for $H_0 : \theta_0 = 0$ using a partition $\mathcal{C}_n$ with G clusters is defined by

$$T_{\mathrm{CCE}(\alpha),n} = \text{ Reject } \quad \text{if } \quad \frac{\widehat{\theta}_{\mathsf{X}_n}}{\widehat{V}_{\mathrm{CCE},\mathcal{C}_n}} > \sqrt{\frac{G}{G-1}} \times t_{1-\alpha/2,G-1}.$$

where $\widehat{V}_{\mathrm{CCE},\mathcal{C}_n}$ is the standard cluster covariance estimator (without a degrees of freedom correction). BCH analyzed the CCE estimator under an asymptotic frame with a fixed, finite number of clusters G (i.e. G independent of n). The resulting inference is based on calculating a t statistic based on an estimated standard error. Under regularity conditions, BCH show that the resulting t statistic is asymptotically pivotal, but distributed according to $\sqrt{G/(G-1)} \times t_{G-1}$ where t_{G-1} is the t distribution with $G-1$ degrees of freedom.

The regularity conditions required in BCH are strong, and in particular, require that the clusters have equal numbers of observations. The k-medoids algorithm does not generally return such a partition. As a result, the CCE estimator is not anticipated to have asymptotically correct size. In the simulation study presented below, in some settings CCE exhibits moderate size distortion relative to the IM and CRS procedures.

2.6.4 Discussion of uniformity

The high-level conditions used for proving Theorems 2.1 and 2.2 involve the assumption of uniform central limit theorems for convergences $\sup_{\mathcal{C}\in\mathscr{C}_n} \|\tilde{S}_{\mathcal{C}} - \tilde{S}^*_{\mathcal{C}}\|_2 \to 0$ for Gaussian random variables $\tilde{S}_{\mathcal{C}}$. Uniform convergence results may be explicitly derived in several cases of interest. First, when each $\mathscr{C}_n$ contains only finitely many partitions $\mathcal{C}$ all of which have cardinality bounded by some $G_{\max}$ independent of n, then uniform convergence follows

immediately from a pointwise convergence result like that presented in Proposition 2.4.

Another cases is to increasing sequences $G_{\max}$. In particular, if $G_{\max}$ increases sufficiently slowly, then uniform analogous of the results Theorems 2.1 and 2.2 may be anticipated by establishing Berry-Esseen-type bounds for dependent processes. Note, for instance, that Jirak (2016) establishes a unidimensional Berry-Esseen bound for sums of weakly dependent random variables. Note also that standard dimension-dependent Berry-Esseen bounds for independent data scale as $O(G^{1/4}n^{-1/2})$ for total variation distance, implying bounds $O(G^{3/4}n^{-1/2})$ when passing to Euclidean distance, i.e., $\|\tilde{S}_{\mathcal{C}} - \tilde{S}^*_{\mathcal{C}}\|_2$. To maintain focus, this paper does not carry out such a program further.

A final case to consider is to fix G but allow $\mathscr{C}_n$ to contain many different partitions. For example, this may be the case encountered by considering the output of k-medoids algorithm with initial starting medoids. In this case, arguments from standard empirical process theory may be used. For brevity, this paper does not carry out such a program further. Note, however, that the set of G-tuples of points in a bounded subset of Euclidean space has known metric entropy properties. These metric entropies may be used to control the metric entropy of tuples in X_n after embedding $\mathsf{X}_n^{1-\varepsilon}$ into a Euclidean space for suitable $\varepsilon > 0$.

Note also, this paper treats the sequence of metric spaces X_n is a fixed (non-random) sequence. As a result, the associated sequence $\mathcal{C}_n$ is non-random given a deterministic clustering procedure. In cases in which the associated metric spaces are random may be handled with an associated random sequence of partitions $\mathcal{C}_n$. Proof of results analogous to Theorems 2.1 and 2.2 could proceed by leveraging the uniform results presented in the appendix, in conjunction with formal definitions and bounds of the complexity of the space of clusterings for classes of distributions on metric spaces X_n.

2.7 Implementation Details

2.7.1 Implementation of k-medoids clustering

This section states the k-medoids algorithm used in the main text. For finite (X, d), let $\mathsf{C} \subseteq \mathsf{X}$ and define the cost of a cluster C with medoid $i \in \mathsf{X}$ to be $\text{cost}(\mathsf{C}, i) = \sum_{j \in \mathsf{C}} d(i,j)^2$. The total cost for a partition $\mathcal{C}$ and set of medoids $\{i_{\mathsf{C}}\}_{\mathsf{C} \in \mathcal{C}}$ is defined by summing over clusters

$$\text{total cost}(\mathcal{C}, \{i_{\mathsf{C}}\}_{\mathsf{C} \in \mathcal{C}}) = \sum_{\mathsf{C} \in \mathcal{C}} \text{cost}(\mathsf{C}, i_{\mathsf{C}}) = \sum_{\mathsf{C} \in \mathcal{C}} \sum_{j \in \mathsf{C}} d(i,j)^2.$$

Algorithm 2.2. *(k-medoids)*

Input. (X, d), G.

Procedure.

Initialize cluster centroids $\{i_1, ..., i_G\} \subset \mathsf{X}_n$ arbitrarily.

While total cost decreases,

(a) For each $k \leq G$, for each $j \notin \{i_1, ..., i_G\}$ compute the cost with new medoids $\{i_1, ..., i_{k-1}, j, i_{k+1}, ...i_G\}$;

(b) Update to new medoids if total cost decreases.

Output. $\mathcal{C}$ *with* $|\mathcal{C}| = G$.

2.7.2 Implementation of cluster-based inference

This section gives additional implementation details for the three inferential procedures, CCE, IM, and CRS, considered in the main text in both the OLS and IV specifications. Specifically, this section gives all omitted details from Steps 1, 2, and 4 of Section 2.3 in the main text.

In order to describe the procedure, introducing vector and matrix notation is helpful. Let Y, X, W, and U be (NT)-row matrices obtained by stacking Y_{de}, X_{de}, $(W'_{de}, 1)$, and

U_{de}, respectively. In the IV model, let Z and V be (NT)-row matrices obtained by stacking Z_{de} and V_{de}, respectively. Let $M_A = I_n - A(A'A)^{-1}A'$ for some matrix A, and for some $B \in \mathbb{R}^{n\times n}$ with rank k, let P_B be some bijective linear transformation from the subspace orthogonal to the one spanned by columns of B (thus $(n-k)$-dimensional), to $\mathbb{R}^{n-k}$.

Step 1: The implementation details in this step differ slightly in the cases of OLS and IV estimation.

– In the OLS model, let $\widehat{U}$ be the vector of residuals from the full-sample least-square estimation. Then, under $U \sim N(0, \Sigma)$,

$$P_{M_{[X,W]}}\widehat{U} \sim N(0, \Sigma_{PMU}),$$

where

$$\Sigma_{PMU} = P_{M_{[X,W]}} M_{[X,W]} \Sigma M_{[X,W]} P'_{M_{[X,W]}}.$$

Note that the covariance matrix Σ_{PMU} is made non-singular by applying the matrix $P_{M_{[X,W]}}$ to $\widehat{U}$. Σ is estimated by QMLE using the model with the parameter $\tau = (\tau_1, \tau_2, \tau_3)'$,

$$\Sigma_{de,d'e'}(\tau) = cov[U_{de}, U_{d'e'}; \tau] = \exp(\tau_1)\exp(-\tau_2^{-1}\|L_d - L_{d'}\|_2 - \tau_3^{-1}|e - e'|), \tag{2.1}$$

which, in the *BASELINE* case, is the correct model. To implement, calculate

$$\widehat{\tau} = \arg\max_{\tau} \left\{ \frac{1}{2}\log\det(\Sigma_{PMU}(\tau)) + \frac{1}{2}\widehat{U}' P'_{M_{[X,W]}} (\Sigma_{PMU}(\tau))^{-1} P_{M_{[X,W]}}\widehat{U} \right\},$$

where

$$\Sigma_{PMU}(\tau) = P_{M_{[X,W]}} M_{[X,W]} \Sigma(\tau) M_{[X,W]} P'_{M_{[X,W]}}$$

and $\Sigma(\tau) = (\Sigma_{de,d'e'}(\tau))_{de,d'e'}$ is the implied covariance matrix of U under τ. The

covariance matrix estimator is thus $\Sigma(\widehat{\tau})$.

– In the IV model, the covariance matrices for the structural and first-stage equations are estimated separately. Let $\widehat{U} = M_W Y - M_W X\widehat{\theta}$ and $\widehat{V} = M_W X - M_W Z\widehat{\pi}$, where $\widehat{\theta}$ is the 2SLS estimator for θ_0 and $\widehat{\pi}$ is the least-square estimator for π. Then, the covariance matrices for U and V are estimated by solving

$$\widehat{\tau}^{\varepsilon} = \arg\max_{\tau} \left\{ \frac{1}{2} \log\det(\Sigma_{PM\varepsilon}(\tau)) + \frac{1}{2} \widehat{\varepsilon}' P'_{M_W} (\Sigma_{PM\varepsilon}(\tau))^{-1} P_{M_W} \widehat{\varepsilon} \right\},$$

where $\Sigma_{PM\varepsilon} = P_{M_W} M_W \Sigma(\tau) M_W P'_{M_W}$, $\Sigma(\tau) = (\Sigma_{de,d'e'}(\tau))_{de,d'e'}$, $\Sigma_{de,d'e'}(\tau)$ is as previously defined, and ε is either U or V. Then, the covariance estimators for U and V are $\Sigma(\widehat{\tau}^U)$ and $\Sigma(\widehat{\tau}^V)$, respectively.

Step 2: This step simulates size and power for all candidate partitions $\mathcal{C} \in \mathscr{C} = \{\mathcal{C}^{(2)}, ..., \mathcal{C}^{(G_{\max})}\}$. Given the covariance estimator(s) from Step 1, simulate independent copies of the observable data for each $* = 1, ..., 1000$ as follows.

– In the OLS model, draw U^* from the distribution $N(0, \Sigma(\widehat{\tau}))$. Reproduce data by

$$Y^*_{de} = \widehat{\alpha} + W'_{de}\widehat{\gamma} + U^*_{de},$$

where $\widehat{\alpha}$ and $\widehat{\gamma}$ are full-sample least-square estimators, and $U^* = (U^*_{de})_{de}$. For each partition and cluster-based inferential procedure, size is calculated under the assumed null $H_0 : \theta_0 = 0$ and power is calculated over varying θ.

– In the IV model, draw (U^*, V^*) such that

$$\begin{pmatrix} U^* \\ V^* \end{pmatrix} \sim N\left(0, \begin{bmatrix} \widehat{\Sigma}_U & \widehat{\rho}\widehat{\Sigma}_U^{1/2}(\widehat{\Sigma}_V^{1/2})' \\ \widehat{\rho}\widehat{\Sigma}_V^{1/2}(\widehat{\Sigma}_U^{1/2})' & \widehat{\Sigma}_V \end{bmatrix}\right),$$

where $\widehat{\Sigma}_U = \Sigma(\widehat{\tau}^U)$, $\widehat{\Sigma}_V = \Sigma(\widehat{\tau}^V)$, and $\widehat{\rho}$ is the empirical correlation between $\widehat{U}$

and $\widehat{V}$. Reproduce data by

$$\begin{cases} Y^*_{de} = \widehat{\alpha} + \theta X^*_{de} + W'_{de}\widehat{\gamma} + U^*_{de} \\ X^*_{de} = \widehat{\mu} + \widehat{\pi} Z_{de} + X'_{de}\widehat{\xi} + V^*_{de} \end{cases},$$

where $\widehat{\alpha}$ and $\widehat{\gamma}$ are full-sample 2SLS estimators, $\widehat{\mu}$, $\widehat{\pi}$, and $\widehat{\xi}$ are full-sample least-square estimators for the first-stage equation, $U^* = (U^*_{de})_{de}$, and $V^* = (V^*_{de})_{de}$. For each partition and a given cluster-based inferential procedure, size is calculated under the assumed null $H_0 : \theta_0 = 0$ and power is calculated over varying alternative values θ.

Step 4: Alternative hypotheses for estimating simulated average power correspond to $\theta \in \{-10/\sqrt{NT}, -9/\sqrt{NT}, \ldots, -1/\sqrt{NT}, 1/\sqrt{NT}, 2/\sqrt{NT}, \ldots, 10/\sqrt{NT}\}$.

Table 2.1: Impact of Early Morning Attacks on Voter Turnout during the 2014 Election

A. Cluster-based Inference									
	$\widehat{\theta}_0$	*s.e.*	*t*-stat	C.I. (sim adj CV)		C.I. (usual CV)		$\widehat{G}$	
LOC	-0.145	0.061	2.385	[-0.308,	0.018]	[-0.265,	-0.025]	205	
CCE	-0.145	0.090	1.609	[-0.382,	0.093]	[-0.358,	0.068]	8	
IM	-0.243	0.254	0.954	[-0.896,	0.411]	[-0.896,	0.411]	6	
CRS	-0.243			[-1.497,	0.084]	[-1.497,	0.084]	6	

B. Estimates							
		Within 6 clusters defined by k-medoids					
	Full	(1)	(2)	(3)	(4)	(5)	(6)
First stage							
Estimate	0.281	-0.020	0.272	0.420	0.432	0.132	0.098
s.e.	0.086	0.214	0.232	0.490	0.180	0.231	0.257
t-stat	3.252	0.092	1.173	0.858	2.402	0.573	0.379
Second stage							
Estimate	-0.145	-0.180	0.054	0.021	0.085	-1.498	0.063
s.e.	0.061	2.384	0.134	0.073	0.071	2.590	0.617
t-stat	2.385	0.075	0.404	0.288	1.196	0.578	0.102

C. Moran tests for spatial dependence
Test for spatial dependence in first time period only: Moran $I = 4.338$, (p-value = 0.00001)
Test for spatial dependence in second time period only: Moran $I = 2.546$, (p-value = 0.011).
Test for spatial dependence in both time periods: Moran $I = 5.387$, (p-value $< 10_{\text{E}}$-06).
Test for inter-temporal dependence: Moran $I = 2.755$, (p-value = 0.006).

Notes: Panel A presents results of the inferential procedures based on selected clusters. Row labels indicate which procedure is used. Column labeled $\widehat{\theta}_0$ reports the IV estimate of θ_0 for the full sample in the rows labeled LOC and CCE and the average of IV estimators of θ_0 of the six clusters in the rows labeled IM and CRS. Column labeled *s.e.* reports the estimated standard errors for each of the procedures. *t*-stat reports *t*-statistic of the IV estimate of θ_0 for each of the procedures. Column C.I (sim adj CV) reports confidence intervals of the IV estimate of θ_0 using the simulation-adjusted critical values for LOC, CCE and IM and using the simulated p-value threshold for CRS. Column C.I (usual CV) reports confidence intervals of the IV estimate of θ_0 using usual critical values. Column $\widehat{G}$ indicates the optimal number of clusters selected in each procedure. Panel B presents IV and first stage estimates of θ_0 and their associated standard errors and *t*-statistic for the full sample and for six groups generated using k-medoids based on geographic distance. Column labeled "Full sample" reproduces the results in Condra et al. (2018). Columns 3–8 display the IV and first stage estimates for each of the 6 sub-groups generated by k-medoids. Panel C presents Moran tests for spatial dependence as described in the main text.

Table 2.2: Simulation Results: OLS

Method	Estim. Bias	Estim. RMSE	Size	Power			
				-1	-0.5	0.5	1
A. BASELINE							
SK	0.011	0.435	0.452	0.930	0.671	0.676	0.931
LOC-U	0.011	0.435	0.687	0.980	0.817	0.816	0.974
LOC	0.011	0.435	0.044	0.682	0.262	0.257	0.662
CCE	0.011	0.435	0.059	0.553	0.209	0.216	0.534
IM	0.004	0.257	0.058	0.925	0.430	0.430	0.906
CRS	0.002	0.260	0.039	0.841	0.378	0.380	0.858
B. SAR							
SK	-0.028	0.858	0.338	0.683	0.498	0.480	0.675
LOC-U	-0.028	0.858	0.638	0.759	0.666	0.670	0.768
LOC	-0.028	0.858	0.307	0.650	0.451	0.463	0.674
CCE	-0.028	0.858	0.037	0.398	0.166	0.171	0.405
IM	-0.007	0.327	0.048	0.728	0.289	0.284	0.734
CRS	-0.006	0.319	0.045	0.728	0.296	0.256	0.739

Notes: Simulation results for estimation in the design described in Section 2.2. The nominal size is 0.05. Estimates are presented for the estimators, SK, LOC-U, LOC, CCE, IM, CRS described in the text. This table displays settings A and B described in the text. For each estimator and settings, columns display method, estimated bias, estimated RMSE, size, and power against 4 alternatives (-1, -0.5, 0.5, 1). Figures are based on 1000 simulation replications.

Table 2.3: Simulation Results: IV

Method	Estim. Median	Estim. MAD	Size	Power			
				-1	-0.5	0.5	1
A. BASELINE							
SK	0.005	0.148	0.426	1.000	0.988	0.823	0.969
LOC-U	0.005	0.148	0.682	0.982	0.892	0.997	1.000
LOC	0.005	0.148	0.076	0.849	0.588	0.655	0.973
CCE	0.005	0.148	0.074	0.802	0.541	0.507	0.924
IM	-0.052	0.106	0.056	0.933	0.693	0.941	0.987
CRS	-0.051	0.104	0.051	0.998	0.979	0.612	0.886
B. SAR							
SK	-0.003	0.279	0.263	0.872	0.600	0.629	0.752
LOC-U	-0.003	0.279	0.488	0.798	0.666	0.785	0.936
LOC	-0.003	0.279	0.248	0.742	0.606	0.551	0.844
CCE	-0.003	0.279	0.071	0.662	0.477	0.338	0.656
IM	-0.063	0.156	0.036	0.682	0.452	0.583	0.820
CRS	-0.089	0.175	0.036	0.820	0.759	0.387	0.661

Notes: Simulation results for estimation in the design described in Section 2.2. The nominal size is 0.05. Estimates are presented for the estimators, SK, LOC-U, LOC, CCE, IM, CRS described in the text. This table displays settings A and B described in the text. For each estimator and settings, columns display method, estimated median, estimated median absolute deviation, size, and power against 4 alternatives (-1, -0.5, 0.5, 1). Figures are based on 1000 simulation replications.

Table 2.4: Clustering: OLS

		G							$\widehat{G}$
		2	3	4	5	6	7	8	
A. BASELINE									
CCE	size	0.037	0.050	0.052	0.053	0.061	0.053	0.049	0.059
	$\widehat{G}$ frequency	0.000	0.000	0.026	0.112	0.189	0.296	0.377	-
IM	size	0.042	0.054	0.064	0.053	0.050	0.051	0.055	0.058
	$\widehat{G}$ frequency	0.000	0.000	0.000	0.004	0.010	0.103	0.883	-
CRS	size	0.000	0.000	0.000	0.000	0.003	0.030	0.039	0.039
	$\widehat{G}$ frequency	0.000	0.000	0.000	0.000	0.013	0.285	0.702	-
B. SAR									
CCE	size	0.046	0.023	0.037	0.048	0.037	0.030	0.034	0.037
	$\widehat{G}$ frequency	0.000	0.001	0.017	0.058	0.145	0.261	0.518	-
IM	size	0.022	0.039	0.037	0.039	0.047	0.050	0.041	0.048
	$\widehat{G}$ frequency	0.000	0.000	0.004	0.281	0.057	0.300	0.358	-
CRS	size	0.000	0.000	0.000	0.000	0.030	0.040	0.041	0.045
	$\widehat{G}$ frequency	0.000	0.000	0.000	0.000	0.003	0.296	0.701	-

Notes: Simulation results for the design described in Section 2.2. Inferential properties are presented are presented for the estimators, CCE, IM, CRS, described in the text. This table displays settings A and B described in the text. Columns under G report results with the number of groups fixed at a certain G. $\widehat{G}$ is the number of clusters chosen by the criterion on size-power tradeoff described in the text. The rows "$\widehat{G}$ frequency" is the frequency of a particular G achieving the highest simulated power among candidate G's in the setting.

Table 2.5: Clustering: IV

		G							$\widehat{G}$
		2	3	4	5	6	7	8	
A. BASELINE									
CCE	size	0.044	0.057	0.054	0.064	0.063	0.071	0.070	0.074
	$\widehat{G}$ frequency	0.000	0.005	0.035	0.076	0.166	0.289	0.429	-
IM	size	0.045	0.058	0.062	0.053	0.057	0.049	0.060	0.056
	$\widehat{G}$ frequency	0.000	0.000	0.002	0.006	0.012	0.081	0.899	-
CRS	size	0.000	0.000	0.000	0.000	0.002	0.034	0.050	0.051
	$\widehat{G}$ frequency	0.000	0.000	0.000	0.000	0.002	0.252	0.746	-
B. SAR									
CCE	size (simulated cv)	0.036	0.042	0.071	0.078	0.064	0.066	0.071	0.071
	$\widehat{G}$ frequency	0.005	0.015	0.050	0.039	0.090	0.215	0.586	-
IM	size	0.027	0.033	0.039	0.038	0.036	0.031	0.025	0.036
	$\widehat{G}$ frequency	0.014	0.038	0.063	0.749	0.034	0.047	0.055	-
CRS	size	0.000	0.000	0.000	0.000	0.026	0.032	0.029	0.036
	$\widehat{G}$ frequency	0.000	0.000	0.000	0.000	0.022	0.393	0.585	-

Notes: Simulation results for the design described in Section 2.2. Inferential properties are presented are presented for the estimators, CCE, IM, CRS, described in the text. This table displays settings A and B described in the text. Columns under G report results with the number of groups fixed at a certain G. $\widehat{G}$ is the number of clusters chosen by the criterion on size-power tradeoff described in the text. The rows "$\widehat{G}$ frequency" is the frequency of a particular G achieving the highest simulated power among candidate G's in the setting.

CHAPTER 3

ESTIMATION AND INFERENCE FOR SYNTHETIC CONTROL METHODS WITH SPILLOVER EFFECTS

3.1 Introduction

The synthetic control method is often used in treatment effect estimation with panel data where only a few units are treated and a small number of post-treatment periods are available. Current estimation and inference procedures for synthetic control methods do not allow for the existence of spillover effects, which are plausible in many applications. This chapter alleviates these concerns by showing that given some knowledge about the spillover effects, it is possible to provide asymptotically unbiased estimators and inference in the presence of spillovers. Our results extend to scenarios with multiple treated units and periods, and cases with stationary or cointegrated factor models.

The synthetic control method (SCM) has gained popularity in empirical studies since its introduction in Abadie and Gardeazabal (2003). When we observe panel data with only a few treated units and post-treatment periods, the SCM can estimate treatment effects. This setting is common in program evaluation, where we often consider state-level policies and have state-level aggregate data. The SCM models the relationship between the treated and untreated units using pre-treatment data. Then the SCM uses the post-treatment data from untreated units to predict the counter-factual values of the treated unit. This process gives us the synthetic control, while the difference between the outcome and predicted counter-factual outcome is the treatment effect estimate. The SCM exploits the pre-treatment data to form counter-factual values with potentially smaller variance, and so in comparative case studies it is often favored over other program evaluation methods such as difference-in-differences. See Abadie and Cattaneo (2018) for review and comparison of econometric methods used in program evaluation.

However, the SCM and its variants assume explicitly or implicitly that untreated units are not affected by the treatment. That is, they rely on the Stable Unit Treatment Value Assumption (SUTVA). This dependence is natural since the SCM uses post-treatment control units to predict the counter-factual values of the treated units, which, however, is not always realistic. For example, when California imposes a cigarette tax, SUTVA implies (among other things) that nobody decides to shift their cigarette purchases to Nevada.

Under SUTVA and a few other regularity conditions within a factor model, treatment effect estimators using a demeaned version of SCM are shown to be inconsistent but asymptotically unbiased by Ferman and Pinto (2019b), even when the pre-treatment fit is imperfect. Unfortunately, in the presence of a spillover effect, this estimator can be severely biased. Intuitively, the reason is that post-treatment controls are contaminated by the spillover effect, resulting in a biased estimator of the counter-factual value of the treated unit in post-treatment periods, which implies a biased treatment effect estimate. Contamination inducing bias is a standard problem in program evaluation, even within difference-in-differences and RCTs. This problem is worse for the SCM. If by chance the spillover is concentrated in control units that the synthetic control method puts significant weight on, the bias will be substantially worse than in difference-in-differences. Moreover, it is possible the spillovers propagate along the same channels as the underlying factor model, which would mean that the SCM may actively select for units which will induce bias. In our simulation section, we will explore this bias in more depth.

It is worth noting that the problem caused by spillover effects cannot be fully solved by naïve methods such as not including contaminated units in estimation. This is because the contaminated units are often the most important control units that can be useful in forming the synthetic control. Simply not including them in estimation can potentially cause efficiency loss. Moreover, there are cases where most or even all control units are affected by the spillover, which cannot be solved by throwing away affected control units.

This is also true for synthetic control methods that are modified to estimate treatment effects with multiple treated units, since the current methods in the literature use only the units that are not affected by the treatment in order to form the synthetic control. For examples of multivariate synthetic control methods, see Cavallo et al. (2013), Firpo and Possebom (2018), Kreif et al. (2016), Robbins et al. (2017), and Xu (2017).

The goal of this chapter is to relax the SUTVA condition and to perform estimation and testing. Particularly, we look at the cases where there are spillover effects, which are defined by a Rubin model as the difference between the actual outcomes and the counterfactual ones. To facilitate estimation, we assume that some knowledge about the spillover effects is known. More specifically, the treatment effect and the spillover effects are linear in some unknown parameters. We give examples where this assumption is plausible. For each unit of observation, we estimate a model between it and all the other units, using the SCM with pre-treatment data. Thanks to the known spillover structure, we obtain asymptotically unbiased estimators for the treatment and spillover effects. We also characterize the asymptotic distribution of the estimator. Unlike the current methods, our method uses information from all control units in estimation.

In addition, we propose an inferential procedure based on Andrews (2003)'s end-of-sample instability test, or P-test. We first generalize the P-test to the synthetic control method without spillover effects and then generalize it further to incorporate cases with spillover effects. Similar to the P-test, our testing procedures use the idea of approximating the null distribution of the statistic using pre-treatment data.

We give high-level conditions under which our methods are valid. Specifically, our conditions adapt to factor models with either stationary or cointegrated common factors, which are often used to justify the usage of synthetic control methods. Furthermore, we consider extensions where treatment applies to multiple units or periods, and where there are extra covariates.

This chapter mainly contributes to three developing literatures. First, it complements the fast-developing literature on synthetic control inference by relaxing SUTVA. Due to its popularity among empirical researchers, many formal results have been developed for statistical inference in similar settings. For example, Conley and Taber (2011) consider hypothesis testing in a similar data structure where only a few units are treated and both pre- and post-treatment periods are short. They consider difference-in-differences, and use control units to form the null distribution of the statistic. In this particular setting with only a few treated units, difference-in-difference estimator can be treated as a special case of the SCM with equal weights. In Ferman and Pinto (2017) and Hahn and Shi (2017), similar ideas are used to conduct placebo tests which permute across observed units. Among all, Chernozhukov et al. (2017) is the most related to our work, since they also use outcomes across periods rather than across units like the above citations. Li (2019) proposes a testing procedure that is based on the idea of projection onto convex sets and results in Fang and Santos (2018). However, none of the papers mentioned above allows for the existence of spillover effects. Our methods provide formal statistical results in this setting, without assuming SUTVA. Furthermore, our estimation and testing procedure applies to factor models with cointegrated common factors, which is of special interest even in cases without spillover effects.

We also contribute to the literature on spillover effects. This fast-growing literature looks into both estimation of treatment effects in the presence of spillover effects, as well as estimation of spillover effects themselves. For example, Vazquez-Bare (2017) consider a framework where observations are grouped into clusters, and spillover effects are allowed within a cluster, but not across clusters. It discusses estimation of heterogeneous treatment effects as a function of the number of treated units within the same cluster, and spillover effects as a function of whether the unit is treated, and number of treated units within the same cluster. Basse et al. (2017) and Rosenbaum (2007) use randomization test for

inference in the presence of spillover effects. Also see Basse et al. (2017) and Vazquez-Bare (2017) for a literature review on spillover effects. However, this literature seldom looks at the panel data setting with only a few treated units and short post-treatment periods. This limitation is in part because we usually do not have enough information about the spillover effects in this particular setting. We overcome this problem by requiring a potentially weak assumption that the spillover structures be pre-specified and follow a pattern that is linear in some underlying parameters. With that specification, we can estimate the spillover effects and perform statistical tests on the spillovers.

Third, our results extend the literature on Andrews (2003)'s end-of-sample instability tests. Andrews (2003) uses data across time periods to approximate the null distribution of the test statistic, and apply this idea to OLS, IV, and GMM. Chernozhukov et al. (2017) propose a permutation test that is more general, but similar in cases where serial correlation matters. We extend this idea to the SCM case, and further to more complicated cases with spillover effects. As Andrews et al. (2006) extends Andrews (2003)'s results to the cointegrated cases, we also show that our method is still valid for a cointegrated factor model.

The remainder of this chapter is organized as follows. Section 3.2 introduces a factor model with spillover effects, proposes an estimator of the spillover effects and derives its asymptotic distribution. Section 3.3 considers the P-test introduced by Andrews (2003) and Andrews et al. (2006), and explains how it can be applied in our settings, with proofs in the Appendix. Section 3.4 extends our methods to cases with multiple treated units and/or multiple post-treatment periods, and briefly discusses cases with extra covariates. Section 3.5 concludes. All proofs are in the appendix.

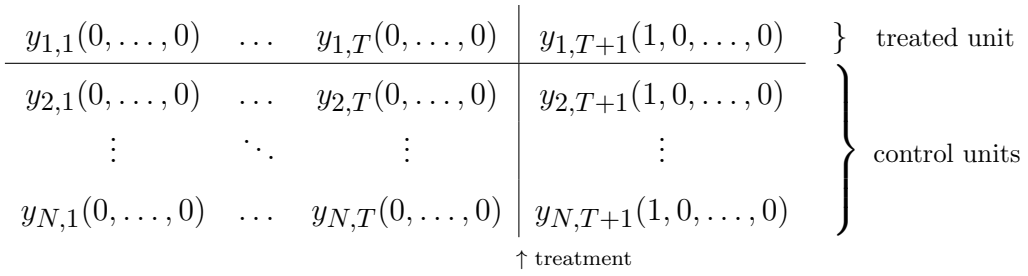


Figure 3.1: Data structure for comparative case studies

3.2 Model and Estimation

3.2.1 A Rubin model with spillover effects

We consider Rubin's potential outcome model. In Rubin's model with violation of SUTVA, the potential outcomes are functions of treatment assignments on all units. Namely, the outcome of unit i at time t is

$$y_{i,t} = y_{i,t}(d_t),$$

where $d_t = (d_{1,t}, \ldots, d_{N,t})'$ and $d_{i,t} = 1$ if unit i has been treated at time t.

We consider a standard synthetic control setting where only one unit is treated and only one period is available after the treatment is implemented. We consider cases with multiple treated units and multiple post-treatment periods in Section 3.4. Let unit 1 be treated between time T and $T+1$, and there be another $N-1$ units that are not directly treated by the policy. Thus, we observe an $N \times (T+1)$ panel as shown in Figure 3.1.

Note that we only observe outcomes with $d_{T+1} = (0, \ldots, 0)'$ or $d_{T+1} = (1, 0, \ldots, 0)'$. This is the fundamental limitation of the dataset we are currently studying. Unless other homogeneity conditions are assumed, we cannot say anything about $y_{i,T+1}(d_{T+1})$ for $d_{T+1} \notin \{(0, \ldots, 0)', (1, 0, \ldots, 0)\}$ because only a few units are treated and only a few post-treatment

periods are available. For notation simplicity, let

$$\begin{cases} y_{i,t}(0) = y_{i,t}(0,\ldots,0) \\ y_{i,t}(1) = y_{i,t}(1,0,\ldots,0) \end{cases}$$

for each (i,t). Let $\alpha_i = y_{i,T+1}(1) - y_{i,T+1}(0)$ be the potential deviation from unit i's counterfactual outcome $y_{i,T+1}(0)$ where no unit is treated at time $T+1$. That is, α_1 is the direct treatment effect on unit 1, while α_i with $i \neq 1$ is the indirect effect or spillover effect. Throughout, we consider the case where N is fixed and T goes to infinity.

In case studies, we are often interested in estimating the treatment effect α_1. For example, Abadie et al. (2010) consider the direct treatment effect on California of the tobacco control policy implemented in the state. A common choice is the synthetic control estimator. Namely, we obtain the synthetic control weights by solving the optimization problem

$$\begin{bmatrix} \widehat{a}_1 \\ \widehat{b}_1 \end{bmatrix} = \underset{\tilde{a}\in\mathbb{R},\tilde{b}\in W^{(1)}}{\arg\min} \sum_{t=1}^{T} (y_{i,t} - \tilde{a} - Y_t'\tilde{b})^2, \tag{3.1}$$

where $Y_t = (y_{1,t},\ldots,y_{N,t})'$ and $W^{(1)} = \{(w_1,\ldots,w_N)' \in \mathbb{R}_+^N : w_1 = 0, \sum_{j=2}^N w_j = 1\}$. An estimator of the treatment effect α_1 is given by

$$\widehat{\alpha}_1 = y_{1,T+1} - (\widehat{a} + Y_{T+1}'\widehat{b}),$$

i.e., the counter-factual value $y_{1,T+1}(0)$ is approximated by $\widehat{a} + Y_{T+1}'\widehat{b}$. For this chapter we use an constraint set as in the demeaned synthetic control method (Ferman and Pinto, 2019b). That is, we do not restrict the intercept but require the other coefficients to be positive and sum up to one. [1]

1. Other choices of constraint set for $(\widehat{a}_1, \widehat{b}_1')'$ include $\{0\} \times \{0\} \times \Delta_{N-1}$ as in the original synthetic control method of Abadie and Gardeazabal (2003) and Abadie et al. (2010), and $\mathbb{R} \times \{0\} \times \mathbb{R}_+^{N-1}$ as in the

3.2.2 Spillover structure

Throughout the chapter, we assume that some knowledge about the spillover effects is known. Namely, assume that the full effect vector α is a linear transformation of some unknown parameter $\gamma \in \mathbb{R}^k$, i.e. $\alpha = A\gamma$. Typically, γ has less dimensions than α does. Here are some examples that fit in this framework.

Example 3.1. *Assume a subset of control units, but not all of them, are equally affected by the spillover effects, i.e.*

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}, \ \gamma = \begin{bmatrix} \alpha_1 \\ b \end{bmatrix}.$$

Example 3.2. *Assume the spillover effect shrinks as the geometric distance goes up. For $i = 2, \ldots, N$, $\alpha_i = b\exp(-d_i)$ where d_i is the distance between unit 1 and unit i and b is some unknown parameter of interest. Then, we have*

$$A = \begin{bmatrix} 1 & 0 \\ 0 & \exp(-d_2) \\ \vdots & \vdots \\ 0 & \exp(-d_N) \end{bmatrix}, \ \gamma = \begin{bmatrix} \alpha_1 \\ b \end{bmatrix}.$$

Example 3.3. *Assume the spillover effect is likely to take place at some known locations, but not at other locations, while the sizes of spillover effects are allowed to vary across those*

---

modified synthetic control of Li (2019), where $\Delta_{N-1} = \{w \in \mathbb{R}^{N-1} : w_i \geq 0 \text{ for each } i, \sum_{i=1}^{N-1} w_i = 1\}$ is a $(N-1)$-dimensional simplex. See Doudchenko and Imbens (2016) for a discussion of other restriction sets.

units. For example, assume there are potential spillovers at locations whose distance to unit 1 *is less than* $\bar{d}$*. Then, the treatment and spillover effect vector can also be represented by* $A\gamma$*. WLOG order the units by increasing distance from unit* 1*, and let* p *the number of units experiencing spillovers. Then*

$$A = \begin{bmatrix} 1 & 0_{1\times p} \\ 0_{p\times 1} & I_p \\ 0_{(N-p-1)\times 1} & 0_{(N-p-1)\times p} \end{bmatrix}, \ \gamma = \begin{bmatrix} \alpha_1 \\ \alpha_{k_1} \\ \vdots \\ \alpha_{k_p} \end{bmatrix}.$$

Thus the units indexed $2, ..., (p+1)$ *each experience their own size spillover effect.*

The assumptions in Example 3.3 are often plausible. If mis-specification of the spillover structure is a concern, one can always choose an A matrix that incorporates more potential spillovers, i.e., a bigger p.

3.2.3 Invertibility assumption

In order to back out the spillover effects, we proceed as follows. We first define the individual synthetic control weights and their limits. Namely, let

$$\begin{bmatrix} \widehat{a}_i \\ \widehat{b}_i \end{bmatrix} = \underset{\tilde{a}\in\mathbb{R},\tilde{b}\in W^{(i)}}{\arg\min} \sum_{t=1}^{T} (y_{i,t} - \tilde{a} - Y_t\tilde{b}')^2, \tag{3.2}$$

where $W^{(i)} = \{(w_1, \ldots, w_N)' \in \mathbb{R}_+^N : w_i = 0, \sum_{j=1}^N w_j = 1\}$. Then, let

$$a_i = \text{plim } \widehat{a}_i, \ b_i = \text{plim } \widehat{b}_i,$$

and we only consider cases where they are well-defined. We show later by Lemma 3.1 that a_i and b_i exist for each i in factor models with stationary or cointegrated common factors.

In general, a_i and b_i do not coincide with the weights that reconstruct the factor loadings (Ferman and Pinto, 2019b).

For each (i,t), define the specification error by

$$u_{i,t} = y_{i,t}(0) - (a_i + Y_t(0)'b_i). \tag{3.3}$$

Note that the i-th entry of b_i is zero. Define $a = (a_1, \ldots, a_N)'$, $B = (b_1, \ldots, b_N)'$, and $M = (I - B)'(I - B)$. Stacking Equation (3.3) for all i's gives

$$u_t = Y_t(0) - (a + BY_t(0)),$$

where and $u_t = (u_{1,t}, \ldots, u_{N,t})'$. For $t = T + 1$, this becomes

$$u_{T+1} = (I - B)(Y_{T+1} - \alpha) - a, \tag{3.4}$$

where $Y_{T+1} = (y_{1,T+1}, \ldots, y_{N,T+1})'$. We will use this equation to estimate the spillover effect.

Defining $M = (I - B)'(I - B)$, we introduce the following invertibility assumption:

Condition IN. *$A'MA$ is non-singular.*

First note Condition IN is testable in principle. We can consistently estimate B so the data informs us of the validity of this assumption. To understand this assumption better, we replace α by $A\gamma$ in Equation (3.4) and have

$$(I - B)A\gamma = (I - B)Y_{T+1} - a - u_{T+1}. \tag{3.5}$$

Equation (3.5) is the key to learning α. Under mild regularity conditions, a and B are identified from the model and learned by the synthetic control method. We do not observe u_{T+1},

but the distribution of u_{T+1} can be learned using pre-treatment data under stationarity of $\{u_t\}_{t\geq 1}$. Therefore, if $A'MA$ is non-singular, or equivalently, $(I-B)A$ has full rank, we can form an estimator of γ whose limiting distribution is identified by multiplying both sides of Equation (3.5) by $(A'MA)^{-1}A'(I-B)'$. Note that we do not identify γ or α. This is because we have only one observation of the outcome in post-treatment periods.

We illustrate Condition IN in the following toy example.

Example 3.4. *Assume there are 3 units in total, where unit 1 is treated. Let the synthetic control weight matrix B be*

$$B = \begin{bmatrix} 0 & w_1 & 1-w_1 \\ w_2 & 0 & 1-w_2 \\ w_3 & 1-w_3 & 0 \end{bmatrix}.$$

Suppose the researcher first assumes unit 2 and 3 are equally exposed to the spillover effects. That is, they assume

$$A_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \ \gamma = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}, \ \textit{and} \ \alpha = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_2 \end{bmatrix}.$$

Then, Condition IN does not hold, because

$$(I-B)A_1 = \begin{bmatrix} 1 & -1 \\ -w_2 & w_2 \\ -w_3 & w_3 \end{bmatrix}.$$

If they instead assumes only one of the controls is exposed to the spillover effects, Condition

IN is satisfied in general. In this case,

$$A_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \gamma = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}, \text{ and } \alpha = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ 0 \end{bmatrix},$$

and

$$(I - B)A_2 = \begin{bmatrix} 1 & -w_1 \\ -w_2 & 1 \\ -w_3 & w_3 - 1 \end{bmatrix}.$$

It can be shown that $(I - B)A_2$ *always has full rank for* $(w_1, w_2, w_3) \in [0, 1]^3$.

This applies to more general settings. That is, if all controls are equally hit by the spillover effects, then $(I - B)A$ does not have full rank and we lose Condition IN. Allowing a few units to be exempt from the spillover effects makes $(I - B)A$ have full rank in general.

A more interesting case is Example 3.3, where we only restrict the range of spillover effects and allow the levels to vary. In this case, $(I - B)A$ can be obtained by eliminating columns that correspond to units that are neither treated nor exposed to spillover effects. Again, as long as at least one control is not exposed to the spillover effects, $(I - B)A$ has full rank in general. This assumption is more convincing if a moderate number of columns are eliminated from $(I - B)$, i.e. only a few units are exposed to the spillover effects.

3.2.4 Estimation

We form estimators for (a, B) using synthetic control methods as in (3.2). We do that for each $i = 1, \ldots, N$, as if each i is the treated unit and other units are controls. Then, the estimators for a and B are $\widehat{a} = (\widehat{a}_1, \ldots, \widehat{a}_N)'$ and $\widehat{B} = (\widehat{b}_1, \ldots, \widehat{b}_N)'$ respectively. Let

$\widehat{M} = (I - \widehat{B})'(I - \widehat{B})$ be an estimator for M. Let an estimator of γ be such that

$$\begin{aligned}\widehat{\gamma} &= \arg\min_{g \in \mathbb{R}^k} \|(I - \widehat{B})(Y_{T+1} - Ag) - \widehat{a}\| \\ &= (A'\widehat{M}A)^{-1}A'(I - \widehat{B})'((I - \widehat{B})Y_{T+1} - \widehat{a}). \end{aligned} \tag{3.6}$$

Note that the FOC implies

$$A'(I - B)'u_{T+1} = 0,$$

i.e. it requires that some weighted sum of the residuals to be zero. Under that condition, the treatment and spillover effect vector α can be estimated by $\widehat{\alpha} = A\widehat{\gamma}$.

Assumption 3.1. (a) $\{u_t\}_{t \geq 1}$ *is stationary, and has mean zero.*

(b) $\|\widehat{a} - a\| = o_p(1)$, $\|\widehat{B} - B\| = o_p(1)$

(c) $\|(\widehat{B} - B)Y_{T+1}(0)\| = o_p(1)$.

(d) $A'MA$ *is non-singular.*

Note that Part (c) excludes polynomial time trends.

Theorem 3.1. *Suppose Assumption 3.1 holds. Then,* $\widehat{\alpha} - (\alpha + Gu_{T+1}) \to_p 0$ *as* $T \to \infty$, *where* $G = A(A'MA)^{-1}A'(I - B)'$. *Moreover,* $E[Gu_{T+1}] = 0$.

The structure of the limiting distribution is similar to the case as in Ferman and Pinto (2019b), as it is inconsistent but asymptotically unbiased (i.e. that the difference between the estimator and the true value has zero mean). Note that consistent estimators are impossible because only one post-treatment period is available.

Moreover, we can form an estimator of α with possibly lower variance. For some positive definite matrix $W \in \mathbb{R}^N$, we minimize $\|W^{1/2}\epsilon_{T+1}\|$ instead of $\|\epsilon_{T+1}\|$. The resulting estimator is

$$\widehat{\gamma}_W = \arg\min_{g \in \mathbb{R}^k} \|W^{1/2}((I - \widehat{B})(Y_{T+1} - Ag) - \widehat{a})\|$$

$$= (A'\widehat{M}_W A)^{-1} A'(I - \widehat{B})'W((I - \widehat{B})Y_{T+1} - \widehat{a}),$$

where $\widehat{M}_W = (I - \widehat{B})'W(I - \widehat{B})$. The corresponding estimator for α is $\widehat{\alpha}_W = A\widehat{\gamma}_W$. In the spirit of GMM with an efficient weighting matrix, let $\Omega = Cov[u_1]$ and W_T^e be a consistent estimator of Ω^{-1}. Then an estimator of α with lower variance can be achieved by $\widehat{\alpha}^e = \widehat{\alpha}_{W_T^e}$.

Let $M_W = (I - B)'W(I - B)$, $G_W = A(A'M_W A)^{-1}A'(I - B)'W$ for some weighting matrix W, $W^e = \Omega^{-1}$, $M^e = M_{W^e}$, and $G^e = G_{W^e}$. Then, we have the following results.

Proposition 3.1. *Suppose Assumption 3.1 holds, W_T is a consistent estimator for W, and W_T^e is a consistent estimator for W^e. Then, $\widehat{\alpha}_{W_T} - (\alpha + G_W u_{T+1}) \to_p 0$, and specifically, $\widehat{\alpha}^e - (\alpha + G^e u_{T+1}) \to_p 0$, as $T \to \infty$. Moreover, $(Cov[G_W u_{T+1}] - Cov[G^e u_{T+1}])$ is positive semi-definite.*

In practice, we need to estimate Ω, and for that we would need relatively large sample size (large T) to have a good approximation.

3.2.5 The factor model as an example

Factor models are often used to justify the usage of synthetic control methods. Here we show that our assumptions are satisfied by factor models with stationary and cointegrated common factors. We follow Ferman and Pinto (2019b) and consider a factor model such that for $i = 1, \ldots, N$ and $t = 1, \ldots, T + 1$,

$$y_{i,t}(0) = \eta_t + \lambda_t' \mu_i + \epsilon_{i,t}, \tag{3.7}$$

where λ_t is F-dimensional common factors, and $\epsilon_{i,t}$ is noise that is uncorrelated with λ_t. For notation simplicity, we write $Y_t(0) = (y_{1,t}(0), \ldots, y_{n,t}(0))'$, $Y_t = (y_{1,t}, \ldots, y_{n,t})'$, and $\epsilon_t = (\epsilon_{1,t}, \ldots, \epsilon_{n,t})'$.

We focus on two sets of conditions in our discussion.

Condition ST (model with stationary common factors)**.** *Assume* $\{(\eta_t, \lambda_t, \epsilon_t)\}_{t\geq 1}$ *is stationary, ergodic for the first and second moments, and has finite* $(2+\delta)$*-moment for some* $\delta > 0$*. Assume* $cov[Y_t(0)] = \Omega_y$ *is positive definite.*

Remark 3.1. *We show in the proof of Lemma 3.1 that in this case*

$$b_i = \underset{w \in W^{(i)}}{\arg\min}\ (w - e_i)'\Omega_y(w - e_i),$$
$$a_i = E[y_{i,1}(0) - Y_1(0)'b_i],$$

where e_i *is a unit vector with one at the* i*-th entry and zeros everywhere else, and* $W^{(i)} = \{(w_1, \ldots, w_N) \in \mathbb{R}_+^N : w_i = 0, \sum_{j\neq i} w_j = 1\}$*. Note that in general* b_i *does not recover the factor structure, because* $\mu_i \neq (\mu_1, \ldots, \mu_N)b_i$ *in general.*

Remark 3.2. *We do not impose any restriction on the factor loadings* $\{\mu_i\}_{i=1}^N$ *except for* Ω_y *being positive definite. In the stationary case, the key for the treatment estimator to be asymptotically unbiased and the test proposed below to be valid is to include an intercept in the optimization problem* (3.2)*.*

Condition CO (model with cointegrated $\mathcal{I}(1)$ common factors)**.** *Rewrite Equation* (3.7) *as*

$$y_{i,t}(0) = (\lambda_t^1)'\mu_i^1 + (\lambda_t^0)'\mu_i^0 + \epsilon_{i,t},$$

and η_t *can be either in* λ_t^1 *or* λ_t^0*. Assume* $\{(\lambda_t^0, \epsilon_t)\}_{t\geq 1}$ *is stationary, ergodic for the first and second moments, and has finite 4-th moment. Without loss of generality,* $E[\epsilon_{i,t}] = 0$*. Assume* $\{\lambda_t^1\}_{t\geq 1}$ *is* $\mathcal{I}(1)$*. Further assume for each* i*,* $y_{i,t}(0)$ *is such that weak convergence holds for* $T^{-1/2}y_{i,[rT]}(0) \Rightarrow \nu_i(r)$*, where* $\Rightarrow$ *is weak convergence and process* $\nu_i(r)$ *is defined on* $[0,1]$ *and has bounded continuous sample path almost surely. For each* i*, let* $W^{(i)} = \{(w_1, \ldots, w_N) \in \mathbb{R}_+^N : w_i = 0, \sum_{j\neq i} w_j = 1\}$*. Assume for each* i*, there exists* $w^{(i)} \in W^{(i)}$ *such that* $\mu_i^1 = \sum_{j=1}^N w_j^{(i)}\mu_j^1$*. That is,* $(w^{(i)} - e_i)$ *is a cointegrating vector for* $Y_t(0)$*, where*

e_i is a unit vector with i-th entry being one and zeros everywhere else.

Note that Condition CO puts restrictions on the factor loadings. The restrictions are similar to those in Ferman and Pinto (2019b).

The relevance of the factor model is given by the following lemma:

Lemma 3.1. *Under Condition IN, either Condition ST or Condition CO implies Assumption 3.1.*

Thus, results derived in Theorem 3.1 apply to factors models with Condition ST or Condition CO.

3.3 Inference

In this section, we discuss formal results on inference. At a high level, our test uses pre-treatment data to form the null distribution of a pre-specified post-treatment quantity. Flexibility in defining that quantity leads to a variety of hypotheses. In Section 3.3.1, we consider the case without spillover effects, and state the assumptions under which Andrews' P test (Andrews, 2003) is valid. In Section 3.3.2, we generalize P test to cases where spillover effects cannot be ignored.

3.3.1 Cases without spillover effects

Suppose for now there are no spillover effects, i.e. $\alpha_2 = \cdots = \alpha_N = 0$. We want to test for the existence of treatment effect on unit 1. The null and alternative hypotheses of interest are

$$\begin{cases} H_0 : \alpha_1 = 0, \\ H_1 : \alpha_1 \neq 0. \end{cases}$$

The test procedure we consider here is the end-of-sample instability test (P-test) in Andrews (2003). The usage of Andrews' test in the context of synthetic control methods is mentioned

in Ferman and Pinto (2019a), where they focus on the difference-in-differences estimator. We formalize this idea and derive conditions under which Andrews' test delivers valid inference results.

We assume the α_1 is independent of T under H_1. That is, we consider fixed, not local, alternatives, as in Andrews (2003) and Andrews et al. (2006). Specifically, α_1 does not change as T grows, which facilitates our analysis of the test statistic under H_1.

Now we translate our hypothesis into the linear formulation considered in Abadie and Gardeazabal (2003). Namely, we have

$$y_t = \begin{cases} a_1 + Y_t' b_1 + u_{1,t}, & \text{for } t = 1, \ldots, T, \\ a_1^* + Y_t' b_1 + u_{1,t}, & \text{for } t = T+1. \end{cases}$$

A non-zero treatment effect is equivalent to a shift in the intercept a_1 (or equivalently, change of the distribution of $u_{1,t}$, at $t = T+1$). The null and alternative hypothesis become

$$\begin{cases} H_0 : a_1^* = a_1, \\ H_1 : a_1^* \neq a_1. \end{cases}$$

Let the synthetic control regression residuals be $\widehat{u}_{1,t} = y_{1,t} - \widehat{a}_1 - Y_t'\widehat{b}_1$. The test statistic is defined by

$$P = \widehat{u}_{1,T+1}^2.$$

For notational simplicity, let $\widehat{\beta}_1 = (\widehat{a}_1, \widehat{b}_1')'$ and $x_t = (1, Y_t')'$. For $\beta \in \mathbb{R}^{N+1}$, define

$$P_t(\beta) = (y_{1,t} - x_t'\beta)^2.$$

Then, $P = (y_{1,T+1} - x_{T+1}'\widehat{\beta}_1)^2 = P_{T+1}(\widehat{\beta}_1)$. Let P_∞ be a random variable with the same

distribution as $P_{T+1}(\beta_1)$ with $\beta_1 = (a_1, b_1')'$. Let $P_t = P_t(\widehat{\beta}_1^{(t)})$, where $\widehat{\beta}_1^{(t)} = \widehat{\beta}_1$ for each t.[2] Define

$$\widehat{F}_{P,T}(x) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{P_t \leq x\},$$

and let $F_P(x)$ be the distribution function of $P_1(\beta_1)$. Finally, let $\widehat{q}_{P,1-\tau} = \inf\{x \in \mathbb{R} : \widehat{F}_{P,T}(x) \geq 1-\tau\}$, and $q_{P,1-\tau}$ be the $(1-\tau)$-quantile of $P_1(\beta_1)$. The assumptions and validity of the testing procedure are established as follows.

Assumption 3.2. (a) $\{u_t\}_{t\geq 1}$ *are stationary, ergodic, and have mean zero.*

(b) $E[|u_t|] < \infty$.

(c) $\exists$ *a non-random sequence of positive definite matrices* $\{C_T\}_{T\geq 1}$ *such that* $\max_{t\leq T+1} \|C_T^{-1} x_t\| = O_p(1)$

(d) $\|C_T(\widehat{\beta}_1 - \beta_1)\| = o_p(1)$, *and* $\max_{t=1,\ldots,T} \|C_T(\widehat{\beta}_1^{(t)} - \beta_1)\| = o_p(1)$.

(e) *The distribution function of* $P_1(\beta_1)$ *is continuous and increasing at its* $(1-\tau)$*-quantile.*

Theorem 3.2. *Suppose Assumption 3.2 holds. Then, as* $T \to \infty$,

(a) $P \to_d P_\infty$ *under* H_0 *and* H_1,

(b) $\widehat{F}_{P,T}(x) \to_p F_P(x)$ *for all* x *in a neighborhood of* $q_{P,1-\tau}$ *under* H_0 *and* H_1,

(c) $\widehat{q}_{P,1-\tau} \to_p q_{P,1-\tau}$ *under* H_0 *and* H_1,

(d) $\Pr(P > \widehat{q}_{P,1-\tau}) \to \tau$ *under* H_0.

In addition, we show the relevance of the factor model in this context by the following lemma:

Lemma 3.2. *Suppose the distribution function of* $P_1(\beta_1)$ *is continuous and increasing at its* $(1-\tau)$*-quantile. Then, either Condition ST or Condition CO implies Assumption 3.2.*

2. Readers can also use leave-one-estimator to construct P_t as in Andrews (2003) and Andrews et al. (2006). For $t = 1, \ldots, T$, the leave-one-out estimator $\widehat{\beta}_1^{(t)}$ is defined by the synthetic control weight estimator using only observations indexed by $s = 1, \ldots, t-1, t+1, \ldots, T$.

3.3.2 Cases with spillover effects

Now we allow for non-zero spillover effects. We propose a testing procedure that is based on Andrews' P-test and accounts for the spillover effect. The null and alternative hypotheses we consider are $H_0 : C\alpha = d$ and $H_1 : C\alpha \neq d$, with C and d known. For example, we want to test for the hypothesis that there is no treatment effect at the treated unit (unit 1), then we let $C = (1, 0, 0, \ldots, 0) \in \mathbb{R}^{1\times N}$ and $d = 0$. This effectively makes Section 3.3.1 a special case of our test, although Theorem 3.2 has slightly stronger results than Theorem 3.3 does. If we want to test that there is a spillover, then we can let $C = [0_{(N-1)\times 1}\ I_{N-1}] \in \mathbb{R}^{(N-1)\times N}$ and $d = (0, \ldots, 0)' \in \mathbb{R}^{(N-1)\times 1}$.

The test statistic we consider here is $P = (C\widehat{\alpha} - d)'W_T(C\widehat{\alpha} - d)$ for some weighting matrix $W_T \to_p W$. Recall $G = A(A'MA)^{-1}A'(I - B)$ and can be consistently estimated by $\widehat{G} = A(A'\widehat{M}A)^{-1}A'(I - \widehat{B})$ if $\widehat{B} \to_p B$. By Theorem 3.1, P is asymptotically equivalent to $u'_{T+1}G'C'WCGu_{T+1}$. To construct critical values, define

$$P_t(\theta) = (Y_t - \theta x_t)'G'C'WCG(Y_t - \theta x_t),$$

and

$$\widehat{P}_t(\theta) = (Y_t - \theta x_t)'\widehat{G}'C'W_TC\widehat{G}(Y_t - \theta x_t),$$

for some $\theta \in \mathbb{R}^{N\times(N+1)}$, $x_t = (1, Y_t')'$, and $\widehat{G} = A(A'\widehat{M}A)^{-1}A'(I - \widehat{B})'$. Let $\widehat{P}_t = \widehat{P}_t(\widehat{\theta}^{(t)})$, where $\widehat{\theta}^{(t)} = \widehat{\theta}$ for each t.[3] Let $P_\infty = P_1(\theta_0)$ for $\theta_0 = [a\ B]$. Define

$$\widehat{F}_{P,T}(x) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{\widehat{P}_t \leq x\},$$

and let $F_P(x)$ be the distribution function of P_∞. Finally, let $\widehat{q}_{P,1-\tau} = \inf\{x \in \mathbb{R} :$

3. Similar to the case without spillover effects, the leave-one-out estimator $\widehat{\theta}^{(t)} = [\widehat{a}^{(t)}\ \widehat{B}^{(t)}]$ is defined by the synthetic control weight estimator using only observations indexed by $s = 1, \ldots, t-1, t+1, \ldots, T$.

$\widehat{F}_{P,T}(x) \geq 1-\tau\}$, and $q_{P,1-\tau}$ be the $(1-\tau)$-quantile of P_∞. The assumptions and validity of the proposed testing procedure are given as follows.

Assumption 3.3. (a) *Assumption 3.1 holds.*

(b) $\{u_t\}_{t\geq 1}$ *is ergodic and* $E[\|u_t\|] < \infty$.

(c) *There exists a non-random sequence of positive definite matrices* $\{D_T\}_{T\geq 1}$ *such that* $\max_{t\leq T+1} \|D_T^{-1}x_t\| = O_p(1)$.

(d) $\|(\widehat{\theta}-\theta_0)D_T\|_F = o_p(1)$, *and* $\max_{t=1,\ldots,T} \|(\widehat{\theta}^{(t)}-\theta_0)D_T\|_F = o_p(1)$, *where* $\|\cdot\|_F$ *is the Frobenius norm.*

(e) *The distribution function of* $P_1(\theta_0)$ *is continuous and increasing at its* $(1-\tau)$*-quantile.*

(f) $W_T \to_p W$ *as* $T \to \infty$.

Theorem 3.3. *Suppose Assumption 3.3 holds. Then, under* H_0, *as* $T \to \infty$,

(a) $P \to_d P_\infty$,

(b) $\widehat{F}_{P,T}(x) \to_p F_P(x)$ *for all* x *in a neighborhood of* $q_{P,1-\tau}$,

(c) $\widehat{q}_{P,1-\tau} \to_p q_{P,1-\tau}$,

(d) $\Pr(P > \widehat{q}_{P,1-\tau}) \to \tau$.

Again, we show the relevance of the factor model in this context by the following lemma:

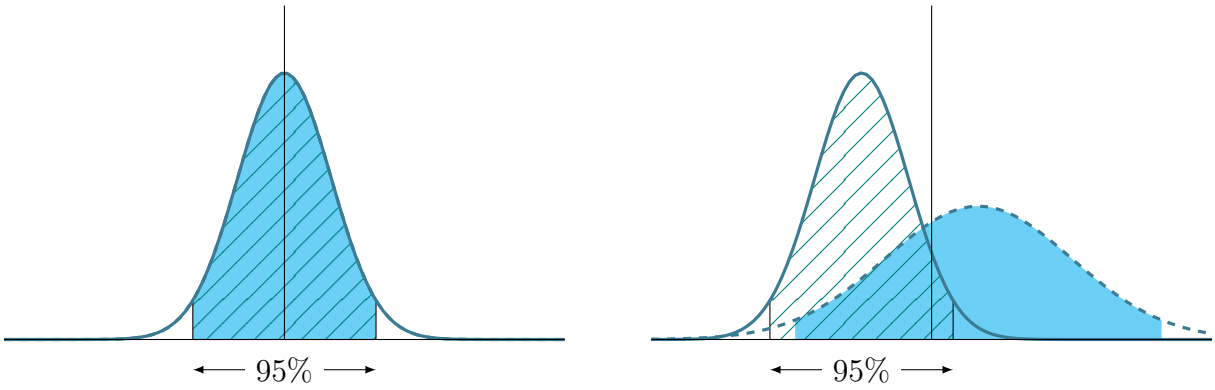
Lemma 3.3. *Suppose the distribution function of* $P_1(\theta_0)$ *is continuous and increasing at its* $(1-\tau)$*-quantile. Then, under Condition IN, Assumption 3.3 is satisfied if either of these holds:*

(i) *Condition ST with* $W_T = I$ *or* $W_T = (C\widehat{G}(T^{-1}\sum_{t=1}^T \widehat{u}_t\widehat{u}_t')\widehat{G}'C')^{-1}$;

(ii) *Condition CO with* $W_T = I$.

3.3.3 Other testing procedures

When we allow for existence of non-zero spillover effects, the existing testing procedures including the placebo test and the original Andrews' test will have poor performance. Here



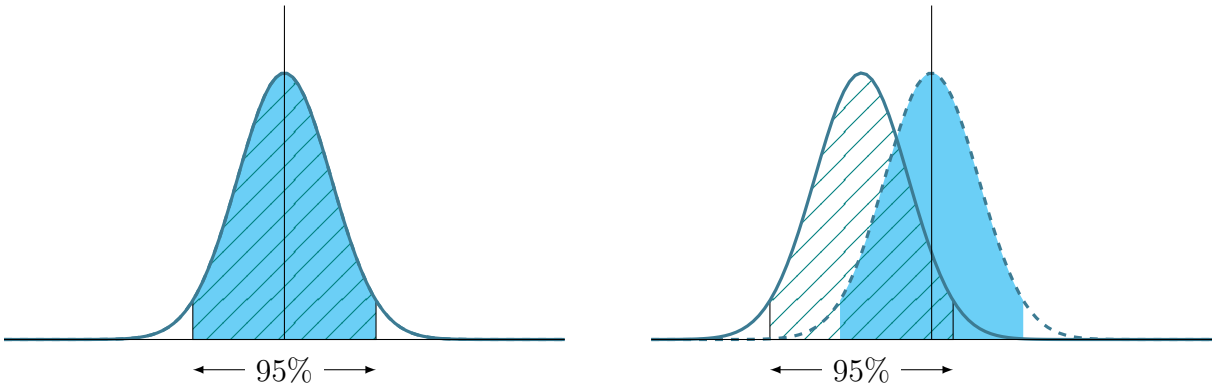
(a) No spillover effects (b) Spillover effects

Figure 3.2: Placebo test with spillover effects

we intuitively explain what happens to placebo test as in Abadie and Gardeazabal (2003) and Andrews' test as in Andrews (2003) in the presence of spillover effects.

Suppose we want to test for the treatment effect being zero and are not aware of the spillover effects. Placebo test and Andrews' test are similar in the sense that they use data to form the null distribution of $u_{1,T+1}$ in order to perform hypothesis testing. The difference is that the placebo test exploits variations of $\{\widehat{u}_{i,T+1}\}_{i=1}^{N}$, while Andrews' test uses variations of $\{\widehat{u}_{1,t}\}_{t=1}^{T+1}$.

We illustrate the placebo test first by Figure 3.2. Area with lines is 95% probability region of the error of the treated unit. Filled area is 95% probability region of null distribution formed in placebo test. A test is rejected when the error of the treated units falls outside of the filled area. When there is no spillover effect, the distribution of $\widehat{u}_{1,T+1}$ and distribution of $\{\widehat{u}_{i,T+1}\}_{i=2}^{N}$ overlap asymptotically. As shown in Figure 3.2(b), when there are positive spillover effects, we will underestimate the treatment effect and the density function of $\widehat{u}_{1,T+1}$ moves to the left. At the same time, some of the control units shift to the right because of the positive spillovers, so density of $\{\widehat{u}_{i,T+1}\}_{i=2}^{N}$ moves to the right and gets wider. In terms of test performance, the shift of $\widehat{u}_{1,T+1}$ is offset by the wider density of $\{\widehat{u}_{i,T+1}\}_{i=2}^{N}$ (harder to reject H_0). In essence, the placebo test becomes much more conservative and has low power.



(a) No spillover effects (b) Spillover effects

Figure 3.3: Andrews' test with spillover effects

Now we consider Andrews' test and illustrate it by Figure 3.3. Again, area with lines is 95% probability region of the error of the treated unit. Filled area is 95% probability region of null distribution formed in Andrews' test. A test is rejected when the error of the treated units falls outside of the filled area. When there is no spillover effect, the distribution of $\widehat{u}_{1,T+1}$ and distribution of $\{\widehat{u}_{1,t}\}_{t=1}^{T}$ overlap asymptotically. As shown in Figure 3.3(b), when there is positive spillover effect, we underestimate the treatment effect and the density function of $\widehat{u}_{1,T+1}$ shifts to the left, while the density of $\{\widehat{u}_{1,t}\}_{t=1}^{T}$ doesn't, since they are pre-treatment and the spillover only happens after the treatment. This results in an invalid test.

3.4 Extensions

3.4.1 Multiple treated units

Our method readily extends to cases where multiple units are treated. In our setting, spillover effects are not distinguished from treatment effects, since one can think of spillover as the treatment on the units that are not directly treated. With a corrected specified structure matrix A, we can perform estimation and testing just as previous sections. For example, suppose $N = 4$, unit 1 and unit 2 are treated, unit 3 is affected by spillover effect, and unit

4 is neither treated nor exposed to spillover effect. Then we can specify

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix},$$

and the resulting estimator $\widehat{\gamma} = (\widehat{\gamma}_1, \widehat{\gamma}_2, \widehat{\gamma}_3)'$ by (3.6) is such that $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ are the treatment effect estimator for unit 1 and unit 2, respectively, and $\widehat{\gamma}_3$ is the spillover effect estimator for unit 3. Tests can be performed accordingly. If one wants to test for the hypothesis that there are no spillover effects, the null is then $H_0 : C\alpha = d$, where $C = (0, 0, 1, 0)$ and $d = 0$.

3.4.2 Multiple post-treatment time periods

Suppose now we have observations of $\{y_{i,t}\}$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T+m$. Treatment is received at $t = T + 1$. The model becomes

$$Y_t = \begin{cases} Y_t(0), \text{ if } t \leq T \\ Y_t(0) + \alpha_t, \text{ otherwise.} \end{cases}$$

Note that we do not allow for spillovers in time. That is, the treatment effect or spillover effects cannot affect future selves. For each $t = T + 1, \ldots, T + m$, we need to specify the spillover structure matrix A_t. Then, an estimator of α_t is

$$\widehat{\alpha}_t = A_t(A_t'\widehat{M}A_t)^{-1}A_t'(I - \widehat{B})'((I - \widehat{B})Y_t - \widehat{a}).$$

That is, we treat $T + s$ period as $T + 1$ and do the same procedure as before. For each $t = T + 1, \ldots, T + m$, we can perform separate tests as introduced in previous sections.

To answer simultaneous questions such as whether there is spillover effect at all, we

can extend the P-test discussed above. Consider the null hypothesis $H_0 : C_t\alpha_t = d_t$ for $t = T+1, \ldots, T+m$. Let $\widehat{P}_t$ be constructed as in Section 3.3.2 for $t = 1, \ldots, T$. For $t = T+1, \ldots, T+m$, let $\widehat{P}_t = (C_t\widehat{\alpha}_t - d_t)'W_T(C_t\widehat{\alpha}_t - d_t)$. We can now form

$$P^{(t)} = \sum_{s=0}^{m-1} \widehat{P}_{t+s}$$

for $t = 1, \ldots, T+1$. The test statistic is then $P^{(T+1)}$, and we use $\{P^{(t)}\}_{t=1}^{T}$ to form its null distribution.

3.4.3 Including covariates

Many empirical researchers are interested including extra covariates when using synthetic control methods. Our framework can be readily adapted to settings with covariates. Suppose we have a vector of observable variables $z_{i,t}$ and want to estimate the treatment effects, while being worried about spillover effects. Following Li (2019), we estimate the least square coefficients for the model

$$y_{i,t}(0) = a_i + \sum_{j \neq i} b_{i,j} y_{j,t}(0) + z_{i,t}'\pi + u_{i,t},$$

with the simplex constraints on $b_{i,j}$ and obtain coefficient estimates $(\widehat{a}_i, \widehat{b}_i, \widehat{\pi}_i)$. This is done for each i. Let $\widehat{g}_t = (z_{1,t}'\widehat{\pi}_i, \ldots, z_{N,t}'\widehat{\pi}_N)'$. Under appropriate regularity conditions, the results of the chapter apply when the intercept estimator $\widehat{a}$ is replaced by $\widehat{a} + \widehat{g}_t$ at time t. For example, the treatment effects estimator now becomes

$$\widehat{\gamma} = (A'\widehat{M}A)^{-1}A'(I - \widehat{B})'((I - \widehat{B})Y_{T+1} - \widehat{a} - \widehat{g}_{T+1}).$$

3.5 Conclusion

The synthetic control method is a powerful tool in treatment effect estimation in the panel data settings, but it does not work in the presence of spillover effects. In this chapter, we relax this assumption and propose an estimation and testing procedure that is robust to the presence of spillover effects. Our method requires specification of the spillover structure, which can be weak (Example 3.3). We derive a set of conditions under which our estimators are asymptotically unbiased. We develop a testing procedure based on Andrews (2003)'s end-of-sample instability tests and show that it is asymptotically unbiased under a set of conditions. We show that our conditions are satisfied by the commonly used factor models, with either stationary or cointegrated common factors. Our methods can be extended to cases with multiple treated units and multiple post-treatment periods, and with extra covariates.

APPENDIX A

APPENDIX FOR CHAPTER 1

A.1 Implementation of the Group-Wise Method with Truncated Unbiased IV Estimator

In this section, I describe the details of implementing the proposed procedure. The idea is simply to replace each quantity by its sample analog. Section A.1.1 discusses the case with only one instrument. Section A.1.2 covers the case with more than one instrument.

A.1.1 One single instrument

The algorithm consists of three steps: group-level estimation, debiasing and truncation, and a t-test.

Step 1 We fix some group g and only use observations in this group. Let the corresponding group-level estimators be $\widehat{\psi}_g = (\widehat{\psi}_{1,g}, \widehat{\psi}_{2,g})'$ as in (1.2), and the residuals be $\{\widehat{U}_i, \widehat{V}_i\}_{i \in I_g}$. Let $\widehat{\Lambda}_g$ be a heteroskedasticity and autocorrelation correction estimator (HAC) of

$$Var\left[\frac{1}{\sqrt{n_g}} \sum_{i \in I_g} \begin{pmatrix} Z_i U_i \\ Z_i V_i \end{pmatrix}\right].$$

In the simulation section, we use the Newey-West estimator with $\lfloor 4(T/100)^{1/4} \rfloor$ lags (Newey and West, 1987). The estimator for $Var[\widehat{\psi}]$ is thus

$$\widehat{\Sigma}_g = \begin{pmatrix} Q_{ZZ,g}^{-1} & 0 \\ 0 & Q_{ZZ,g}^{-1} \end{pmatrix} \widehat{\Lambda}_g \begin{pmatrix} Q_{ZZ,g}^{-1} & 0 \\ 0 & Q_{ZZ,g}^{-1} \end{pmatrix},$$

where $Q_{ZZ,g} = n_g^{-1} \sum_{i \in I_g} Z_i Z_i'$. The group-level $(\widehat{\delta}, \widehat{\tau})$ is given by

$$\widehat{\delta}_g = \widehat{\delta}(\widehat{\psi}_g, \widehat{\Sigma}_g), \ \widehat{\tau}_g = \widehat{\tau}(\widehat{\psi}_g, \widehat{\Sigma}_g),$$

and the unbiased IV is $\widehat{\beta}_g = \widehat{\delta}\widehat{\tau} + \widehat{\sigma}_{12,g}/\widehat{\sigma}_{2,g}^2$.

Step 2 Consider a uniform truncation parameter where $\pi_g^* = \pi^*$ for each g. Let

$$\pi_{SIV}^* = \min_g \frac{1}{\sqrt{n_g}} \Psi^{-1}\left(c\sqrt{\frac{\bar{n}}{n_g}}\right)$$

and

$$\pi_{WIV}^* = \min_g \frac{1}{\sqrt{n_g}} \Psi^{-1}\left(\sqrt{\frac{\underline{n}}{n_g}} \Psi\left(-c\sqrt{\frac{n}{\bar{n}}}\right)\right),$$

where $\bar{n} = \max_g n_g$ and $\underline{n} = \min_g n_g$. The former is suggested by assumptions under the strong IV asymptotics in Section 1.3.2 and the latter by the weak IV asymptotics in Section 1.3.3. The truncation parameter is chosen to be $\pi^* = \min\{\pi_{SIV}^*, \pi_{WIV}^*\}$. In practice, I recommend using $c = 10$. Using the selected threshold π^*, we can obtain a set of group-level truncated unbiased IV estimators $\{\tilde{\beta}_g\}_{g=1}^G$.

Step 3 We apply the t-test to the set of group-level estimators $\{\tilde{\beta}_g\}_{g=1}^G$. Namely, let

$$t = \frac{\bar{\beta} - \beta_0}{\text{se}},$$

where

$$\bar{\beta} = \frac{1}{G} \sum_{g=1}^G \tilde{\beta}_g$$

and

$$\text{se} = \sqrt{\frac{1}{G(G-1)} \sum_{g=1}^G (\tilde{\beta}_g - \bar{\beta})^2}.$$

We reject the null hypothesis $H_0 : \beta = \beta_0$ if $|t| > cv$, where cv is the $(1-\alpha)$-quantile of the t-distribution of $G-1$ degrees of freedom.

A.1.2 Multiple instruments

When multiple instruments are available, we follow Andrews and Armstrong (2017) and use a weighted average of unbiased IV estimators with respect to all instruments. Namely, for the j-th instrument, we perform Steps 1 and 2 as in section A.1.1 and obtain the j-th unbiased IV estimator $\tilde{\beta}_{g,j}$. The group-level estimator is then given by

$$\tilde{\beta}_g = \sum_j w_j \tilde{\beta}_{g,j},$$

where $\{w_j\}_{j=1}^k$ is a set of weights that sum up to one. See Andrews and Armstrong (2017) for a discussion on optimal weight selection. In the simulation section, $\{w_j\}_{j=1}^k$ are simply chosen to be equal weights. Finally, we follow Step 3 in section A.1.1 using $\{\tilde{\beta}_g\}_{g=1}^G$.

A.2 Truncation Parameter Choices

In this section, we investigate the impact of the choice of the truncation parameter. As in Appendix A.1, we recommend using $\pi^* = \min\{\pi^*_{SIV}, \pi^*_{WIV}\}$ with $c = 10$ as the truncation parameter. We look into different choices of c in this experiment.

The data regenerating process is the same as in Section 1.4.1. FMUT methods with three different values of c are reported. The CCE method is also reported for comparison. The power curves are shown in Figure A.1. Generally, the proposed method is quite robust to the choice of the truncation parameter in terms of null rejection rate. Moreover, Figure A.1 exhibits a "bias-variance" tradeoff. That is, a smaller c corresponds to high power but causes more bias.

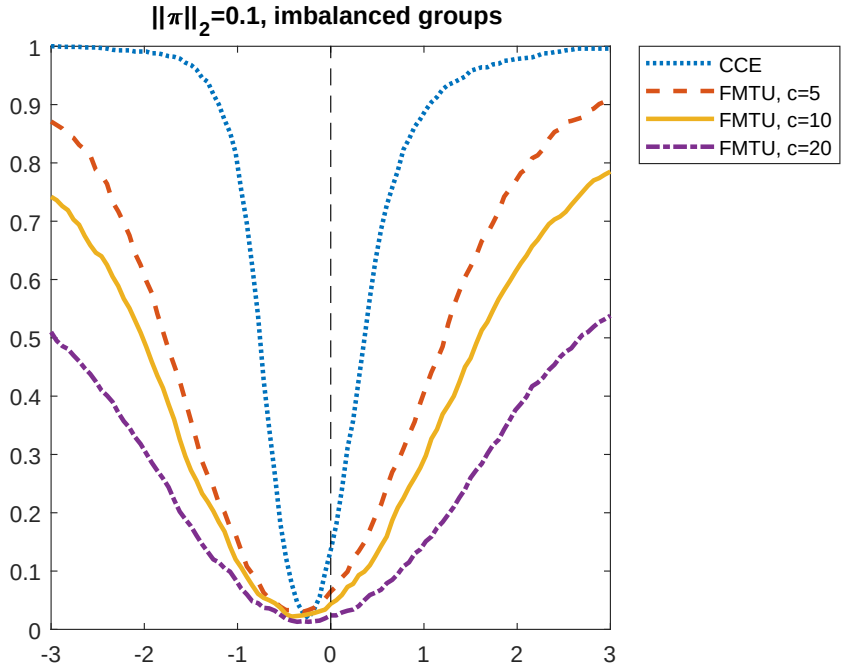


Figure A.1: Power comparison among truncation-parameter choices ($\alpha = 0.05$)

A.3 Useful Results

In this section, I present some results that are useful for proofs in Appendix A.4.

Lemma A.1. $E[\widehat{\tau}\mathbb{1}\{\widehat{\pi} \geq \pi^*\}] = \eta\pi^{-1}$, *where*

$$\eta = (1 - \Phi((\pi^* - \pi)/\sigma_2)) - (1 - \Phi(\pi^*/\sigma_2))\exp(\pi\pi^*/\sigma_2^2 - \pi^2/(2\sigma_2^2)).$$

Proof.

$$\begin{aligned} & E[\widehat{\tau}\mathbb{1}\{\widehat{\pi} \geq \pi^*\}] \\ = \; & E\left[\frac{1}{\sigma_2} \cdot \frac{1 - \Phi(\widehat{\pi}/\sigma_2)}{\phi(\widehat{\pi}/\sigma_2)}\mathbb{1}\{\widehat{\pi} \geq \pi^*\}\right] \\ = \; & \int_{\pi^*/\sigma_2}^{\infty} \frac{1}{\sigma_2} \cdot \frac{1 - \Phi(x)}{\phi(x)}\phi(x - \pi/\sigma_2)\mathrm{d}x \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sigma_2}\int_{\pi^*/\sigma_2}^{\infty}(1-\Phi(x))\exp(x\pi/\sigma_2-\pi^2/(2\sigma_2^2))\mathrm{d}x \\
&= \frac{1}{\pi}\exp(-\pi^2/(2\sigma_2^2))\left((1-\Phi(x))\exp\left(\frac{\pi}{\sigma_2}x\right)\Big|_{\pi^*/\sigma_2}^{\infty}-\int_{\pi^*/\sigma_2}^{\infty}\exp(x\pi/\sigma_2)\mathrm{d}(1-\Phi(x))\right) \\
&= \frac{\eta}{\pi}.
\end{aligned}
$$

The fourth equality is integration by parts. □

Kummer's confluent hypergeometric functions

$$K(a,b,z)=\sum_{k=0}^{\infty}\frac{a^{\bar{k}}}{b^{\bar{k}}}\frac{z^k}{k!},$$

where the rising factorial is defined by

$$x^{\bar{k}}=x(x+1)\ldots(x+n-1).$$

For $z<0$,

$$K(a,b,z)=\frac{\Gamma(b)}{\Gamma(b-a)}(-z)^{-a}[1+O(|z|^{-1})],$$

where the Gamma function is

$$\Gamma(x)=\int_0^{\infty}u^{x-1}e^{-u}\mathrm{d}u.$$

See Abramowitz and Stegun (1965) for reference.

A.4 Proofs

Proof of Proposition 1.1. Let η be defined as in Lemma A.1. Note

$$|E[\tilde{\beta}]-E[\widehat{\beta}_U]| \quad = \quad |E[\widehat{\delta}(\tilde{\tau}-\widehat{\tau})]|$$

$$
\begin{aligned}
&= |E[\widehat{\delta}]| \cdot |E[\tilde{\tau} - \widehat{\tau}]| \\
&= |E[\widehat{\delta}]| \cdot E[(\widehat{\tau} - \tau^*)\mathbb{1}\{\widehat{\pi} < \pi^*\}] \\
&\leq |E[\widehat{\delta}]| \cdot E[\widehat{\tau}\mathbb{1}\{\widehat{\pi} < \pi^*\}] \\
&= |\pi(\beta - \sigma_{12}/\sigma_2^2)| \cdot (\pi^{-1} - \eta\pi^{-1}) \\
&= |\beta - \sigma_{12}/\sigma_2^2| \cdot (1 - \eta) \\
&\to 0.
\end{aligned}
$$

The second equality uses the independence between $\widehat{\delta}$ and $\widehat{\pi}$, which implies the independence between $\widehat{\delta}$ and functions of $\widehat{\pi}$. The inequality is because Ψ is strictly decreasing, and thus $0 \leq \widehat{\tau} - \tau^* \leq \widehat{\tau}$ under the event $\widehat{\pi} < \pi^*$. The fourth equation is by Lemma A.1 and the fact that $E[\widehat{\tau}] = 1/\pi$. The convergence is because $\eta \to 1$ under (i), (ii), and (iii), and σ_{12}/σ_2^2 is bounded. □

Proof of Theorem 1.1. By Assumption 1.1 and following the proof of Proposition 1.1,

$$\sup_g |E[\tilde{\beta}_g] - \beta| \to 0.$$

So

$$t = t^* + \frac{G^{-1}\sum_{g=1}^G (E[\tilde{\beta}_g] - \beta)}{\text{se}} = t^* + o_p(1),$$

where

$$t^* = \frac{G^{-1}\sum_{g=1}^G (\tilde{\beta}_g - E[\tilde{\beta}_g])}{\text{se}}.$$

Under Assumption 1.2, for some absolute constant C,

$$
\begin{aligned}
\sup_x |\Pr(t^* < x) - \Phi(x)| &\leq CB^{-3}\sum_{g=1}^G E[|\tilde{\beta}_g - \beta|^3] \\
&\lesssim CB^{-3}G\max_g E[|\widehat{\delta}_g|^3]E[|\tilde{\tau}_g|^3] \\
&\lesssim CB^{-3}G\bar{\sigma}_\delta^3\kappa M^3
\end{aligned}
$$

$$= \quad o(1). \tag{A.1}$$

The inequality is by a Berry-Esseen bound for Student's statistic in Bentkus et al. (1996). The third line uses a representation of the third raw absolute moment of normal distribution (e.g., see Winkelbauer, 2012). Combining (A.1) with $t = t^* + o_p(1)$, we obtain $t \xrightarrow{d} N(0,1)$.

□

Proof of Proposition 1.2. Assumption 1.1(i) holds automatically by Assumption S1(ii). For 1(ii), note

$$\max_g \frac{\pi_g^* - \pi}{\sigma_{2,g}} = \max_g \Psi^{-1}(\sigma_{2,g} M) - \frac{\pi}{\sigma_{2,g}} \leq \Psi^{-1}(\underline{\sigma}_2 M) \to -\infty$$

by S1(iii) and the fact that $\Psi^{-1}(x) \to -\infty$ as $x \to \infty$. For 1(iii), note

$$\max_g \frac{\pi \pi_g^*}{\sigma_{2,g}^2} = \max_g \frac{\pi \Psi^{-1}(\sigma_{2,g} M)}{\sigma_{2,g}} \leq \frac{\pi \Psi^{-1}(\underline{\sigma}_2 M)}{\bar{\sigma}_2} \to -\infty$$

by S1(iii).

To see Assumption 1.2, first note for some constant C_1, C_2,

$$K\left(-\frac{3}{2}, \frac{1}{2}; -\frac{\mu_\delta^2}{2\sigma_\delta^2}\right) \leq C_1 \left(\frac{\mu_\delta^2}{2\sigma_\delta^2}\right)^{3/2} + C_2 \left(\frac{\mu_\delta^2}{2\sigma_\delta^2}\right)^{1/2} \lesssim C_1 \underline{\sigma}_\delta^{-3}$$

by properties of Kummer's confluent hypergeometric function (e.g., 13.1.5 of Abramowitz and Stegun, 1965). Therefore,

$$M = o\left(\frac{B}{\underline{\sigma}_2 (\bar{\sigma}_2^{-3} G)^{1/3}}\right) = o\left(\frac{B}{\underline{\sigma}_2 (\kappa G)^{1/3}}\right).$$

□

Proof of Proposition 1.3. Assumptions 1.1(i) and (ii) follow the same reasoning as in

the proof of Proposition 1.2. For 1(iii), note

$$\max_g \frac{\pi\pi_g^*}{\sigma_{2,g}^2} \lesssim \frac{\Psi^{-1}(\underline{\sigma}_2 M)}{\sqrt{n}\bar{\sigma}_2} \to -\infty.$$

To see Assumption 1.2, for some constant C,

$$K\left(-\frac{3}{2}, \frac{1}{2}; -\frac{\mu_\delta^2}{2\sigma_\delta^2}\right) \leq 1 + C\left|-\frac{\mu_\delta^2}{2\sigma_\delta^2}\right| = 1 + O\left(\frac{\pi^2}{\underline{\sigma}_\delta^2}\right), \tag{A.2}$$

by properties of Kummer's confluent hypergeometric function. Combining (A.2) with W2(ii) gives Assumption 1.2. □

APPENDIX B

APPENDIX FOR CHAPTER 2

This supplement contains proofs as well as additional simulation results as described in the text.

B.1 Proofs of Propositions 2.1–2.3

Before proving those propositions, the basic structural results about sequences of uniformly Ahlfors spaces are gathered. The first preliminary result is due to Assoud (1977) and ensures the existence of certain embedding of metric spaces of bounded doubling into Euclidean spaces.

Theorem B.1 (Assoud)**.** *Let* (X, d) *be an arbitrary (not necessarily finite) metric space such that the doubling dimension* $K = dim_{2\times}(\mathsf{X}) < \infty$*. Let* $\varepsilon \in (0, 1)$*. Then there exists an* L*-bi-Lipschitz map* $(\mathsf{X}, \mathrm{d}^{\varepsilon}) \to \mathbb{R}^r$ *for some* L, ν *which depend only on* ε *and* K*.*

Ahlfors regularity of X implies constant doubling. As a result, the Proposition 2.5 holds by standard arguments. Proposition B.1 is the relevant corollary for the contexts of Propositions 1 and 2 in the main text.

Proposition B.1. *Suppose that* X *satisfies* C, δ*-finite-Ahlfors regularity. Then the doubling dimension* $dim_{2\times}(\mathsf{X})$ *is bounded by* $dim_{2\times}(\mathsf{X}) \leq \delta \log_2(3C^2)$*.*

Proposition B.2. *Let* (X, d) *be* C, δ*-regular. Then* $(\mathsf{X}, \mathrm{d}^{3/4})$ *has an* L*-bi-Lipschitz map into* $\mathbb{R}^r$ *where* r *and the Lipschitz constant* L *depend only on* C, δ*.*

B.1.1 Proof of Propositions 2.1

In the case that X_n do not embed isometrically, let L be the bi-Lipschitz constant from the maps $(\mathsf{X}_n, \mathrm{d}_n^{1-1/4}) \to \tilde{\mathsf{X}}_n \subseteq \mathbb{R}^{\nu}$ where ν can be taken $\nu = g(C, \delta)$ for the function g defined

prior to the statement of Assumption 2.2. The array $\{\{\zeta_i\}_{i\in \mathsf{X}_n}\}_{n=1}^{\infty}$ indexed on X_n yields an array $\{\{\tilde{\zeta}_i\}_{i\in \tilde{\mathsf{X}}_n}\}_{n=1}^{\infty}$ indexed on $\tilde{\mathsf{X}}_n$. It is sufficient to check the conditions of Corollary 1 in Jenish and Prucha (2009) for this new process. Apply the same set array of constants c_i to $\tilde{\zeta}_i$. Assumption 1 in Jenish and Prucha (2009) is satisfied by the fact that distances are at least ρ_0 in X_n for some $\rho_0 > 0$ by Ahlfors regularity. Note that L depends only on C, δ and in particular, does not change with n. Then $\forall i, j$, $\tilde{\mathrm{d}}_n(i,j)$ is also bounded away from zero by a constant which does not depend on n. Assumption 2.2(ii) is identical to Equation 3 in Jenish and Prucha (2009).

The next conditions in Jenish and Prucha (2009) are mixing conditions. To verify these, let $\tilde{\alpha}_{k,l,n}(r)$ and $\bar{\tilde{\alpha}}_{k,l,n}(r)$ denote the corresponding mixing coefficients for $\tilde{\zeta}_i$ over $\tilde{\mathsf{X}}_n$. Note that $\tilde{\mathrm{d}}(\mathsf{U},\mathsf{V}) \geq r \Rightarrow \mathrm{d}(\mathsf{U},\mathsf{V}) \geq L^{-1}\nu^{-1/2}r^{\frac{3}{4}}$. Let $c = L^{-1}\nu^{-1/2}$. Then $\bar{\tilde{\alpha}}_{k,l}(r) \leq \bar{\alpha}_{k,l}(cr^{3/4})$. To verify Equation 4 in JP, it is sufficient to show that

$$\sum_{m=1}^{\infty} \bar{\tilde{\alpha}}_{1,1}(m) m^{\nu\times\frac{\mu+2}{\mu}-1} < \infty.$$

Note that $\bar{\tilde{\alpha}}_{1,1}(m)$ is nonincreasing and defined for nonnegative real m and $m^{\nu\times\frac{\mu+2}{\mu}-1}$ is a polynomial in m. Thus, the above summation is bounded up to a constant by the corresponding integral

$$\int_{m=0}^{\infty} \bar{\tilde{\alpha}}_{1,1}(m) m^{\nu\times\frac{\mu+2}{\mu}-1} dm$$

Substituting $\bar{\tilde{\alpha}}_{1,1}(m) \leq \bar{\alpha}_{1,1}(cm^{3/4})$ and a standard calculus change of variables $m' = cm^{3/4}$, $dm' = \frac{3}{4}cm^{-1/4}dm$ shows that it is sufficient to verify

$$\int_{m'=0}^{\infty} \bar{\alpha}_{1,1}(m') {m'}^{\frac{4}{3}(\nu\times\frac{\mu+2}{\mu}-1)} {m'}^{1/3} dm' < \infty.$$

This integral is in turn bounded by a constant times the summation

$$\sum_{m'=1}^{\infty} \bar{\alpha}_{1,1}(m') m'^{\frac{4}{3}\nu \times \frac{\mu+2}{\mu}-1} = \sum_{m'=1}^{\infty} \bar{\alpha}_{1,1}(m') m'^{\nu \times \frac{\mu+2}{\mu}-1} < \infty$$

which is assumed to be finite under Assumption 2.2(iii), thus verifying Equation 4 in Jenish and Prucha (2009). Using similar arguments, Assumption 2.2(iv) implies Assumption 4(2) in Jenish and Prucha (2009). Next, Assumption 2.2(v) implies that

$$\bar{\tilde{\alpha}}_{1,\infty}(m) \leq \bar{\alpha}_{1,\infty}(cm^{3/4}) = O((cm^{3/4})^{\frac{4}{3}(\nu-\mu)}) = O(m^{\nu-\mu})$$

thus verifying Assumption 4(3) in JP. Finally, Assumption 2.2(vi) implies Assumption 5 in Jenish and Prucha (2009). This verifies the assumptions of Corollary 1 in Jenish and Prucha (2009). The Almost-Sure Representation theorem (Theorem 2.19 in van der Vaart (1998)) then gives a random variables $\tilde{S}^*$ and $\tilde{S}_{\mathcal{C}}$ such that $\lim_{n\to\infty} \|\tilde{S}^* - \tilde{S}_{\mathcal{C}}\|_2 \to 0$. Setting $\tilde{S}^*_{\mathcal{C}} = \tilde{S}^*$ for all $\mathcal{C}$ gives the result.

B.1.2 Proof of Propositions 2.2

Proposition 2.2 is proven in the case $|\mathcal{C}| = 2$. In this case for $\mathsf{C} \subseteq \mathsf{X}_n$, take $\mathsf{D} = \mathsf{X}_n \setminus \mathsf{C}$. The general case follows by applying the same arguments, and by retracing boundaries when necessary. By arguments in Bester et al. (2011b)(which as noted in the main text were previously also present in Jenish and Prucha (2009) and Bolthausen (1982)), $\sigma(\mathsf{C})^{-1}\sigma(\mathsf{D})^{-1}n = O(1)$ and

$$\left|\operatorname{cov}\left(\sigma(\mathsf{C})^{-1}\sum_{i\in\mathsf{C}}\zeta_j, \sigma(\mathsf{D})^{-1}\sum_{j\in\mathsf{D}}\zeta_j\right)\right| \leq \sigma(\mathsf{C})^{-1}\sigma(\mathsf{D})^{-1}\sum_{(i,j)\in\mathsf{C}\times\mathsf{D}} \bar{\alpha}_{1,1}(\lceil \mathrm{d}_n(i,j)\rceil)^{\mu/(2+\mu)}$$

$$= \sigma(\mathsf{C})^{-1}\sigma(\mathsf{D})^{-1} \sum_{k=1}^{\infty} |\{(i,j) \in \mathsf{C} \times \mathsf{D} : k-1 \leq \mathrm{d}_n(i,j) < k\}| \bar{\alpha}_{1,1}(k)^{\mu/(2+\mu)}.$$

Note that

$$|\{(i,j) \in \mathsf{C} \times \mathsf{D} : k-1 \leq \mathrm{d}_n(i,j) < k\}| \leq |\{(i,j) \in \mathsf{X} \times \mathsf{X} : k-1 \leq \mathrm{d}_n(i,j) < k\}| \leq nCk^{\delta}$$

by Assumption 2.1. By Assumption 2.3 it also follows that

$$\max_{k \leq r} |\{(i,j) \in \mathsf{C} \times \mathsf{D} : k-1 \leq \mathrm{d}_n(i,j) < k\}| \leq o(n).$$

Then the original covariance is bounded by

$$\sigma(\mathsf{C})^{-1}\sigma(\mathsf{D})^{-1} \left[\sum_{k=1}^{r} o(n)\bar{\alpha}_{1,1}(k) + \sum_{m=r}^{\infty} nCk^{\delta}\bar{\alpha}_{1,1}(k) \right]$$

The first term satisfies $\sigma(\mathsf{C})^{-1}\sigma_{\mathsf{D}}^{-1} \sum_{k=1}^{r} o(n)\bar{\alpha}_{1,1}(k) \to 0$ by

$$\sum_{k=1}^{r} o(n)\bar{\alpha}_{1,1}(k) \leq \sum_{k=1}^{\infty} o(n)\bar{\alpha}_{1,1}(k) < \infty$$

and $\sigma_{\mathsf{C}}^{-1}\sigma(\mathsf{D})^{-1}o(n) \to 0$. The second term satisfies $\sigma_{\mathsf{C}}^{-1}\sigma_{\mathsf{D}}^{-1} \sum_{m=r}^{\infty} nCk^{\delta}\bar{\alpha}_{1,1}(k) \to 0$ by $n\sigma(\mathsf{C})^{-1}\sigma(\mathsf{D})^{-1} = O(1)$, $\sum_{m=1}^{\infty} nCk^{\delta}\bar{\alpha}_{1,1}(k) < \infty$ and $r \to \infty \implies \sum_{m=r}^{\infty} Ck^{\delta}\bar{\alpha}_{1,1}(k) \to 0$. The Cramèr-Wold device (over a common space $\mathbb{R}^G$ and linear maps ℓ_G) may applied using a composition of linear maps $\ell_{\mathcal{C}_n} : \mathbb{R}^{\mathcal{C}_n} \xrightarrow{\iota_{\mathcal{C}_n}} \mathbb{R}^G \xrightarrow{\ell_G} \mathbb{R}$ where $\iota_{\mathcal{C}_n}$ are inclusions. Applying the Almost-Sure Representation theorem (Theorem 2.19 in van der Vaart (1998)) in the same way as in the proof of Proposition 2.1 then gives the result.

B.1.3 Proof of Propositions 2.3

First for simplicity consider the case in which G is independent of n. Let $\bar{r}_n = \log n$. When possible, from this point on, n is excluded from notation. Consider two points $i, j \in \mathsf{X}$. Let $M = \{l : |\mathrm{d}(i,l) - \mathrm{d}(j,l)| \leq \bar{r}\}$. Let M_0 be an $\bar{r}^2$-net of M. Suppose for sake of contradiction that $|M| \neq o(n)$. Then $|M_0| \neq o(n/\bar{r}^{2\delta})$. Let $A \subseteq [\frac{1}{2}, \frac{9}{10}]$ satisfy $|a - a'| \geq \bar{r}^3/\mathrm{d}(i,j)$ for each $a, a' \in A$. Take $|A| \geq \frac{4}{10}\frac{\mathrm{d}(i,j)}{2}/\bar{r}^3$. For $a \in A$, let M_a consist of interpolants l_a such that $|\mathrm{d}(i,l_a) - a\mathrm{d}(i,l)| \leq K$ and $|\mathrm{d}(l_a,l) - (1-a)\mathrm{d}(i,l)| \leq K$ for each $l \in M_0$. Then by trigonometry, $\cup_{a\in A} M_a$ is a $3\bar{r}$-separated for n sufficiently large and contains $|A| \times |M_0|$ elements. This can be checked in more detail, by constructing the line segments $\overline{\iota(i)\iota(l)}$ and $\overline{\iota(i)\iota(l')}$ where ι is the coarse isometry to Euclidean space E. Then there are points u, u' which belong to the above constructed line segments with distances $d_{\mathsf{E}}(\iota(l_a), u)$, $d_{\mathsf{E}}(\iota(l'_a), u')$ bounded. u, u' are then shown sufficiently separated to yield the claim. As a result,

$$\left| \bigcup_{a\in A, l_a \in M_a} \mathbf{B}_{\bar{r}}(l_a) \right| \geq |A| \times |M_0| C^{-1} \bar{r}^{\delta} \geq \frac{4}{20}\frac{\mathrm{d}(i,j)}{\bar{r}^3} |M_0| C^{-1} \bar{r}^{\delta}.$$

But by $|M_0| \neq o(n/\bar{r}^{2\delta})$, it follows that the above quantity must be larger than n infinitely often provided $\mathrm{d}(i,j) \geq \bar{r}^{4+\delta}$ for n sufficiently large. This is impossible. Therefore, the small boundaries result is shown once it is shown that k-medoids terminates with medoids $i_1, ..., i_k$ such that $\mathrm{d}(i_k, i_l) \geq (\log n)^{4+\delta}$ for n sufficiently large. Again for contradiction, suppose there is a sequence $\ell_n = o(1)$ such that for infinitely many n, there are two clusters $\mathsf{C}_1, \mathsf{C}_2$ with medoids i_1, i_2 satisfying $\mathrm{d}(i_1, i_2) < \ell_n n^{1/\delta}$. By the pigeonhole principal, there must be a cluster C_3 with n/G members. Then $diam(\mathsf{C}_3)$ must be at least $C^{-1}(n/G)^{1/\delta}$. Let x_3 be the corresponding medoid. Then there must be $i'_3 \in \mathsf{C}_3$ such that $\mathrm{d}(i_3, i'_3) \geq \frac{1}{4}C^{-1}(n/G)^{1/\delta}$ and $\mathrm{d}(i'_3, i_k) \geq \frac{1}{4}C^{-1}(n/G)^{1/\delta}$ for any other medoid i_k. Then consider the update in the partitioned medoid algorithm given by $i_2 \leftarrow i'_3$. This update is cost reducing for n sufficiently large. To see this, note that for elements, $i \in \mathbf{B}_{\frac{1}{4}C^{-1}(n/G)^{1/\delta}}(i'_3)$ the total cost reduction

from being reassigned from a medoid centered around i_2 to a medoid centered around i_3' is at least $|\mathbf{B}_{\frac{1}{4}C^{-1}(n/G)^{1/\delta}}(i_3')|\frac{1}{4}C^{-1}(n/G)^{1/\delta} \geq \frac{1}{4}C^{-1}(n/G)^{1/\delta}C^{-1}(\frac{1}{4}C^{-1}(n/G)^{1/\delta})^{\delta}$. The total cost increase from reassigning elements in C_2 to C_1 is at most $\ell_n n^{1/\delta}|\mathsf{C}_2| \leq \ell_n n^{1/\delta} n$. The difference between the above two quantities is a lower bound on the cost reduction for the update. Comparing the above to quantities for n sufficiently large, the k-medoids algorithm could not have stopped at a step with $\mathrm{d}(i_1, i_2) < \ell_n n^{1/\delta}$ giving the desired contradiction. Finally, note that $\mathrm{d}(i_k, i_l) \geq \ell_n n^{1/\delta}$ for all medoids i_k, i_l, some ℓ_n bounded uniformly away from 0, and for n sufficiently large implies the balanced clusters condition after applying Ahlfors regularity. For G increasing with $G = o(n)$ the same argument holds with $\check{n} := \lfloor n/G \rfloor$ replacing n in the appropriate places.

B.2 Proof of Theorem 2.1

Define the function ψ by $\psi(S) = \mathbf{1}_{|t(S)|>t_{1-\alpha/2,G-1}}$, where S has G groups. Let $\tilde{\mathrm{E}}$ be expectation with respect to the measure $\tilde{\Pr}$. Note that for a partition $\mathcal{C}$, $T_{\mathrm{IM}(\alpha),\mathcal{C}} = \psi(S_{\mathcal{C}})$. By construction of $\tilde{S}^*_{\mathcal{C}}$ and Theorem 1 of Ibragimov and Müller (2010), for any n and any $\mathcal{C} \in \mathscr{C}_n$,

$$\tilde{\mathrm{E}}[\psi(\tilde{S}^*_{\mathcal{C}})] \leq \alpha.$$

Note that $(\mathrm{E}[\psi(S_{\widehat{\mathcal{C}}})] - \alpha)_+ \leq |\mathrm{E}[\psi(S_{\widehat{\mathcal{C}}})] - \tilde{\mathrm{E}}[\psi(S^*_{\mathcal{C}})]| + (\tilde{\mathrm{E}}[\psi(S^*_{\mathcal{C}})] - \alpha)_+$ which is equal to $|\tilde{\mathrm{E}}[\psi(\tilde{S}_{\mathcal{C}})] - \tilde{\mathrm{E}}[\psi(S^*_{\mathcal{C}})]|$. Next, to show $\sup_{\mathcal{C}\in\mathscr{C}_n} |\tilde{\mathrm{E}}[\psi(\tilde{S}_{\mathcal{C}})] - \tilde{\mathrm{E}}[\psi(\tilde{S}^*_{\mathcal{C}})]| \to 0$, define the events

$$\mathcal{E}_{1,\mathcal{C}} = \{|t(\tilde{S}^*_{\mathcal{C}}) - t_{1-\alpha/2,|\mathcal{C}|-1}| > b_{\mathcal{C}}\}, \quad \mathcal{E}_{2,\mathcal{C}} = \{|\mathrm{sd}(\tilde{S}^*_{\mathcal{C}})| > b'_{\mathcal{C}}\}$$

for some $b_{\mathcal{C}}, b'_{\mathcal{C}} \in \mathbb{R}$. Note that by the upper and lower bounds on the variances of $\tilde{S}^*_{\mathcal{C}}$ in Assumption 2.6, $t(\tilde{S}^*_{\mathcal{C}})$ and $\mathrm{sd}(\tilde{S}^*_{\mathcal{C}})$ are all random variables with well-defined density which are all upper bounded everywhere by some common constant, denoted M, which is independent of $\mathcal{C}$. Thus, $\Pr(\mathcal{E}_{1,\mathcal{C}} \cap \mathcal{E}_{2,\mathcal{C}}) \geq 1 - M(b_{\mathcal{C}} + b'_{\mathcal{C}})$. Then on $\mathcal{E}_{1,\mathcal{C}}$, a sufficient

condition for $\psi(\tilde{S}_{\mathcal{C}}) = \psi(\tilde{S}^*_{\mathcal{C}})$ is that $|t(\tilde{S}_{\mathcal{C}}) - t(\tilde{S}^*_{\mathcal{C}})| < b_{\mathcal{C}}$. By $\sup_{\mathcal{C}\in\mathscr{C}_n} \|\tilde{S}^*_{\mathcal{C}} - \tilde{S}_{\mathcal{C}}\|_2 \to 0$, and by continuity properties of $t(\cdot)$ outside of $\mathcal{E}_{2,\mathcal{C}}$, there are $b_{\mathcal{C}} < b_n \to 0, b'_{\mathcal{C}} < b'_n \to 0$ such that $|t(\tilde{S}_{\mathcal{C}}) - t(\tilde{S}^*_{\mathcal{C}})| < b_{\mathcal{C}}$ is satisfied on $\mathcal{E}_{1,\mathcal{C}} \cap \mathcal{E}_{2,\mathcal{C}}$. Therefore, $|\tilde{\mathrm{E}}[\psi(\tilde{S}_{\mathcal{C}})] - \tilde{\mathrm{E}}[\psi(\tilde{S}^*_{\mathcal{C}})]| \leq 1 - M(b_n + b'_n)$, for all $\mathcal{C} \in \mathscr{C}_n$ from which the uniform result follows. Finally,

$$\begin{aligned}\Pr(\widehat{T}_{\mathrm{IM}(\alpha),n} = \text{ Reject }) &= \Pr(T_{\mathrm{IM}(\widehat{\alpha}),\widehat{\mathcal{C}}} = \text{ Reject }) \leq \Pr(T_{\mathrm{IM}(\alpha),\widehat{\mathcal{C}}} = \text{ Reject }) \\ &= \sum_{\mathcal{C}\in\mathscr{C}_n} \Pr(T_{\mathrm{IM}(\alpha),\mathcal{C}} = \text{ Reject } | \widehat{\mathcal{C}} = \mathcal{C}) \Pr(\widehat{\mathcal{C}} = \mathcal{C})\end{aligned}$$

Bound the above expression by $o(1) + \sum_{\mathcal{C}\in\mathscr{C}_{0,n}} \Pr(T_{\mathrm{IM}(\alpha),\mathcal{C}} = \text{ Reject }) \Pr(\widehat{\mathcal{C}} = \mathcal{C}) \leq o(1) + \alpha$ by using both Assumption 2.4(i) and (ii). This concludes the proof of Theorem 2.1.

B.3 Proof of Theorem 2.2

Define functions $\phi(S)$ and $\tilde{\phi}(S, U)$ for $S \in \mathbb{R}^{\mathcal{C}}$ and $U \in \mathbb{R}$ according to

$$\phi(S) = \begin{cases} 1, & \text{if } w(S) > w^{(k)}(S) \\ a(S), & \text{if } w(S) = w^{(k)}(S) \\ 0, & \text{if } w(S) < w^{(k)}(S) \end{cases}$$

and define $\tilde{\phi}(S, U) = \phi(S)$ if $W(S) \neq W^{(k)}(S)$ and $\mathbf{1}_{U<a(S)}$ otherwise.

By Theorem 2.1 in Canay et al. (2017), using the fact that $h(\tilde{S}^*_{\mathcal{C}}) =_d \tilde{S}^*_{\mathcal{C}}$ for all h by the fact that h acts as component-wise multiplication by signs, it follows that $\mathrm{E}[\phi(\tilde{S}^*_{\mathcal{C}})] = \alpha$. Note,

$$\mathrm{E}[\phi(S_{\mathcal{C}})] - \alpha = \tilde{\mathrm{E}}[\tilde{\phi}(\tilde{S}_{\mathcal{C}}, \tilde{U}) - \tilde{\phi}(\tilde{S}^*_{\mathcal{C}}, \tilde{U})].$$

Therefore, it is sufficient to show $\limsup_{n\to\infty} \sup_{\mathcal{C}\in\mathscr{C}_n} \tilde{\mathrm{E}}[\tilde{\phi}(\tilde{S}_{\mathcal{C}}, \tilde{U}) - \tilde{\phi}(\tilde{S}^*_{\mathcal{C}}, \tilde{U})] = 0$.

Consider an integer $\check{G}$, and decompose $\mathscr{C}_n = \mathscr{C}_{n,\check{G}} \cup \mathscr{C}^c_{n,\check{G}}$ where $\mathscr{C}_{n,\check{G}} = \{\mathcal{C} \in \mathscr{C}_n : |\mathcal{C}| \leq$

$\check{G}\}$ and $\mathscr{C}^c_{n,\check{G}} = \{\mathcal{C} \in \mathscr{C}_n : |\mathcal{C}| > \check{G}\}$, which are handled separately. For each $\mathcal{C} \in \mathscr{C}_{n,\check{G}}$, let $\mathcal{E}_\mathcal{C}$ be the event in which the order statistics of $\{w(h\tilde{S}_\mathcal{C}) : h \in \mathcal{H}_\mathcal{C}\}$ and $\{w(h\tilde{S}^*_\mathcal{C}) : h \in \mathcal{H}_\mathcal{C}\}$ correspond to the same transformations $h^{(1)}, \ldots, h^{(|\mathcal{H}_\mathcal{C}|)}$. Then the previous expression can be controlled by

$$\begin{aligned}|\tilde{\mathrm{E}}[\tilde{\phi}(\tilde{S}_\mathcal{C}, \tilde{U}) - \tilde{\phi}(\tilde{S}^*_\mathcal{C}, \tilde{U})]| =& |\tilde{\mathrm{E}}[\tilde{\phi}(\tilde{S}_\mathcal{C}, \tilde{U})\mathbf{1}_{\mathcal{E}_\mathcal{C}} + \tilde{\phi}(\tilde{S}_\mathcal{C}, \tilde{U})\mathbf{1}_{\mathcal{E}^c_\mathcal{C}} - \tilde{\phi}(\tilde{S}^*_\mathcal{C}, \tilde{U})\mathbf{1}_{\mathcal{E}_\mathcal{C}} - \tilde{\phi}(\tilde{S}^*_\mathcal{C}, \tilde{U})\mathbf{1}_{\mathcal{E}^c_\mathcal{C}}]| \\ =& |\tilde{\mathrm{E}}[\tilde{\phi}(\tilde{S}_\mathcal{C}, \tilde{U})\mathbf{1}_{\mathcal{E}^c_\mathcal{C}} - \tilde{\phi}(\tilde{S}^*_\mathcal{C}, \tilde{U})\mathbf{1}_{\mathcal{E}^c_\mathcal{C}}]| \\ =& |\tilde{\mathrm{E}}[(\tilde{\phi}(\tilde{S}_\mathcal{C}, \tilde{U}) - \tilde{\phi}(\tilde{S}^*_\mathcal{C}, \tilde{U}))\mathbf{1}_{\mathcal{E}^c_\mathcal{C}}]| \leq 2\tilde{\mathrm{E}}[\mathbf{1}_{\mathcal{E}^c_\mathcal{C}}].\end{aligned}$$

Furthermore, $\sup_{\mathcal{C} \in \mathscr{C}_{n,\check{G}}} \tilde{\mathrm{E}}[\mathbf{1}_{\mathcal{E}^c_\mathcal{C}}]$ can be controlled using Fatou's Lemma by

$$\limsup_{n\to\infty} \sup_{\mathcal{C} \in \mathscr{C}_{n,\check{G}}} \tilde{\mathrm{E}}[\mathbf{1}_{\mathcal{E}^c_\mathcal{C}}] \leq \limsup_{n\to\infty} \tilde{\mathrm{E}}[\sup_{\mathcal{C} \in \mathscr{C}_{n,\check{G}}} \mathbf{1}_{\mathcal{E}^c_\mathcal{C}}] \leq \tilde{\mathrm{E}}[\limsup_{n\to\infty} \sup_{\mathcal{C} \in \mathscr{C}_{n,\check{G}}} \mathbf{1}_{\mathcal{E}^c_\mathcal{C}}].$$

To further bound the right-hand side of the above expression, consider any fixed $\omega \in \tilde{\Omega}$ which satisfies the conditions $\sup_{\mathcal{C} \in \mathscr{C}_n} \|\tilde{S}_\mathcal{C}(\omega) - \tilde{S}^*_\mathcal{C}(\omega)\|_2 \to 0$ and

$$\inf_{\mathcal{C} \in \mathscr{C}_{n,\check{G}}} \inf_{h \nsim h' \in \mathcal{H}_\mathcal{C}} |w_\mathcal{C}(h(\tilde{S}^*_\mathcal{C})(\omega)) - w_\mathcal{C}(h'(\tilde{S}^*_\mathcal{C})(\omega))| > \delta_\omega > 0$$

for some $\delta_\omega > 0$. Here, the equivalence $\sim$ is defined by $h \sim h'$ whenever $w \circ h = w \circ h'$. By Assumption 2.6, the set of such $\omega \in \tilde{\Omega}$ has $\tilde{\mathrm{P}}$r-measure 1. Note, more explicitly, choosing such $\delta_\omega > 0$ is possible due to (1) the continuity of the function w and the continuity of the action of h outside a set of $\tilde{P}$-measure 0, and (2) the fact that $|\mathscr{C}_{n,\check{G}}|$ depends only on $\check{G}$. Let $h^{(1)}_\mathcal{C}(\omega), \ldots, h^{(|\mathcal{H}_\mathcal{C}|)}_\mathcal{C}(\omega)$ be such that $w(h^{(1)}_\mathcal{C}(\omega)\tilde{S}^*_\mathcal{C}(\omega)) \leq \cdots \leq w(h^{(|\mathcal{H}_\mathcal{C}|)}_\mathcal{C}(\omega)\tilde{S}^*_\mathcal{C}(\omega))$. Then the expression

$$\begin{aligned}w(h^{(j+1)}_{n,\mathcal{C}}(\omega)\tilde{S}_\mathcal{C}(\omega)) - w(h^{(j)}_{n,\mathcal{C}}(\omega)\tilde{S}_\mathcal{C}(\omega)) =& \; w(h^{(j+1)}_{n,\mathcal{C}}(\omega)\tilde{S}_\mathcal{C}(\omega)) - w(h^{(j+1)}_{n,\mathcal{C}}(\omega)\tilde{S}^*_\mathcal{C}(\omega)) \\ &+ w(h^{(j+1)}_{n,\mathcal{C}}(\omega)\tilde{S}^*_{n,\mathcal{C}}(\omega)) - w(h^{(j)}_{n,\mathcal{C}}(\omega)\tilde{S}^*_{n,\mathcal{C}}(\omega))\end{aligned}$$

$$+ w(h^{(j)}_{n,\mathcal{C}}(\omega)\tilde{S}^*_{n,\mathcal{C}}(\omega)) - w(h^{(j)}_{n,\mathcal{C}}(\omega)\tilde{S}_{\mathcal{C}}(\omega))$$

is nonnegative and is furthermore strictly positive for at least one j unless $h \sim h'$ for all $h, h' \in \mathcal{H}_{\mathcal{C}}$ in which case Theorem 2.2 holds trivially. Positivity of the above expression in the case of non-equivalent-under-$\sim$ action follows from noting that the first and the third terms are smaller than $\delta_\omega/2$ in absolute value for n sufficiently large, while the second term is greater than δ_ω.

The claim $\limsup_{n\to\infty} \sup_{\mathcal{C}\in\mathscr{C}_n} \mathbf{1}_{\mathcal{E}^c_{\mathcal{C}}}(\omega) = 0$ $\tilde{\Pr}$-a.s. follows. Thus,

$$\limsup_{n\to\infty} \sup_{\mathcal{C}\in\mathscr{C}_{n,\check{G}}} \tilde{\mathrm{E}}[\mathbf{1}_{\mathcal{E}^c_{\mathcal{C}}}] = 0.$$

Therefore, there is a sequence $\check{G}_n \to \infty$ sufficiently slowly such that

$$\limsup_{n\to\infty} \sup_{\mathcal{C}\in\mathscr{C}_{n,\check{G}_n}} \tilde{\mathrm{E}}[\mathbf{1}_{\mathcal{E}^c_{\mathcal{C}}}] = 0.$$

Next consider the partitions in the complement$\mathcal{C} \in \mathscr{C}^c_{\check{G}_n}$. Let $\mathcal{A}_s \subseteq \mathbb{R}^{\mathcal{C}}$ be of the form $\mathcal{A}_S = \{hS : h \in \mathcal{H}_{\mathcal{C}}\}$ for some $S \in \mathbb{R}^{\mathcal{C}}$. Then w^k is constant over $\mathcal{A}_S$ and thus $w^k(\mathcal{A}_S)$ is well defined. Then for any $\delta > 0$,

$$\tilde{\Pr}(|w^k(\mathcal{A}_S) - w(\tilde{S}^*_{\mathcal{C}})| < \delta | \tilde{S}^*_{\mathcal{C}} \in \mathcal{A}_S) \le \sup_{r\in\mathbb{R}} \tilde{\Pr}(|r - w(\tilde{S}^*_{\mathcal{C}})| < \delta | \tilde{S}^*_{\mathcal{C}} \in \mathcal{A}_S).$$

Let $\Pr_{\mathcal{C}}$ be the uniform probability over $h \in \mathcal{H}_{\mathcal{C}}$. Then the above expression is bounded by

$$p_{\mathcal{C},S}(\delta) := \sup_{r\in\mathbb{R}} \Pr_{\mathcal{C}}(|r - w(hS)| < \delta)$$

Next, for any $\mathcal{A}_{\mathcal{C}}$ which is a disjoint union of sets of the form $\mathcal{A}_S$. Then

$$\tilde{\Pr}(|w^k(\tilde{S}^*_{\mathcal{C}}) - w(\tilde{S}^*_{\mathcal{C}})| \geq \delta) \geq \tilde{\Pr}(|w^k(\tilde{S}^*_{\mathcal{C}}) - w(\tilde{S}^*_{\mathcal{C}})| \geq \delta | \tilde{S}^*_{\mathcal{C}} \in \mathcal{A}_{\mathcal{C}}) \tilde{\Pr}(\mathcal{A}_{\mathcal{C}})$$

$$\geq (1 - \sup p_{\mathcal{C},S}(\delta)) \tilde{\Pr}(\mathcal{A}_{\mathcal{C}})$$

where the supremum runs over S indexing $\mathcal{A}_S \subseteq \mathcal{A}_{\mathcal{C}}$.

Note that a sufficient condition for $\tilde{\phi}(\tilde{S}^*_{\mathcal{C}}, \tilde{U}) = \tilde{\phi}(\tilde{S}_{\mathcal{C}}, \tilde{U})$ is that $\text{sign}(w^k(\tilde{S}^*_{\mathcal{C}}) - w(\tilde{S}^*_{\mathcal{C}})) = \text{sign}(w^k(\tilde{S}_{\mathcal{C}}) - w(\tilde{S}_{\mathcal{C}}))$. A further sufficient condition for this is that for some $\delta > 0$, $|w^k(\tilde{S}^*_{\mathcal{C}}) - w(\tilde{S}^*_{\mathcal{C}})| \geq \delta$, $|w^k(\tilde{S}^*_{\mathcal{C}}) - w^k(\tilde{S}_{\mathcal{C}})| < \delta/2$, and $|w(\tilde{S}^*_{\mathcal{C}}) - w(\tilde{S}_{\mathcal{C}})| < \delta/2$. By Assumption 2.6(iii), w has Lipschitz constant 1 with respect to Euclidean norm on $\mathbb{R}^{\mathcal{C}}$ on a suitably chosen sequence of events of probability approaching 1. Then $|w^k(\tilde{S}^*_{\mathcal{C}}) - w^k(\tilde{S}_{\mathcal{C}})| < \delta/2$ whenever $\|\tilde{S}^*_{\mathcal{C}} - \tilde{S}_{\mathcal{C}}\|_2 < \delta/2$ in which case it also holds that $|w(\tilde{S}^*_{\mathcal{C}}) - w(\tilde{S}_{\mathcal{C}})| < \delta/2$. By Assumption 2.6, a sequence $\check{\delta}_n$ may be chosen such that $\check{\delta}_n \to 0$ sufficiently slowly such that for n sufficiently large, each of the above three inequalities may be achieved with common bounds on an event with probability $1 - o_{\tilde{\Pr}}(1)$.

Finally, using the same reasoning as at the conclusion of Theorem 2.1, invoking Assumption 2.4, $\Pr(\widehat{T}_{\text{CRS}(\alpha),n} = \text{ Reject }) = o(1) + \sum_{\mathcal{C} \in \mathscr{C}_{0,n}} \Pr(T_{\text{CRS}(\alpha),\mathcal{C}} = \text{ Reject }) \Pr(\widehat{\mathcal{C}} = \mathcal{C}) \leq o(1) + \alpha$. concluding the proof of Theorem 2.2.

B.4 Proof of Proposition 2.5

Fix $S \in \mathbb{R}^{\mathcal{C}}$. Let D be a constant chosen later, and consider a rescaled version of w defined by $w(S) = |\mathcal{C}|^{-1/2}|\bar{S}|/(D\text{sd}(S))$. Next bound the following anticoncentration quantity:

$$\sup_{r \in \mathbb{R}} \Pr_{\mathcal{C}}(w(hS) \in [r, r + \delta]).$$

Because $\mathcal{H}_{\mathcal{C}}$ is finite, the range of the supremum above can be restricted and the above

quantity is further reduced by

$$= \sup_{r \in [\min_{h\in\mathcal{H}_\mathcal{C}} w(hS), \max_{h\in\mathcal{H}_\mathcal{C}} w(hS)]} \Pr_{\mathcal{C}}(w(hS) \in [r, r+\delta])$$

$$\leq \sup_{r \in [\min_{h\in\mathcal{H}_\mathcal{C}} w(hS), \max_{h\in\mathcal{H}_\mathcal{C}} w(hS)]} \Pr_{\mathcal{C}}(|\bar{hS}| \in [\min_{h\in\mathcal{H}_\mathcal{C}} aD\mathrm{sd}(hS), \max_{h\in\mathcal{H}_\mathcal{C}} (a+\delta)D\mathrm{sd}(hS)])$$

Note that the length of the interval inside the $\Pr_{\mathcal{C}}$ never exceeds

$$\ell = \ell(\delta, D, S) := \sup_{r \in [\min_{h\in\mathcal{H}_\mathcal{C}} w(hS), \max_{h\in\mathcal{H}_\mathcal{C}} w(hS)]} \max_{h\in\mathcal{H}_\mathcal{C}} (r+\delta)D\mathrm{sd}(hS) - \min_{h\in\mathcal{H}_\mathcal{C}} rD\mathrm{sd}(hS).$$

Therefore,

$$\sup_{r\in\mathbb{R}} \Pr_{\mathcal{C}}(w(hS) \in [r, r+\delta]) \leq \sup_{r\in\mathbb{R}} \Pr_{\mathcal{C}}(|\mathcal{C}|^{-1/2}|\bar{hS}| \in [r, r+\ell]).$$

By considering both branches of the absolute value function, (which gives an additional factor of 2), and expressing the above quantity more explicitly in terms of components h_{C}, it follows that

$$\sup_{r\in\mathbb{R}} \Pr_{\mathcal{C}}(|r - |\mathcal{C}|^{-1/2}| \sum_{\mathsf{C}\in\mathcal{C}} h_{\mathsf{C}} S_{\mathcal{C}})|| < \ell) \leq 2 \sup_{r\in\mathbb{R}} \Pr_{\mathcal{C}}(|r - |\mathcal{C}|^{-1/2} \sum_{\mathsf{C}\in\mathcal{C}} h_{\mathsf{C}} S_{\mathcal{C}})| < \ell).$$

By corollary 2.9 in Rudelson and Vershynin (2008), which states an anticoncentration bound for bernoulli sums, this is bounded by

$$2 \sup_{r\in\mathbb{R}} \Pr_{\mathcal{C}}(|r - |\mathcal{C}|^{-1/2} \sum_{\mathsf{C}\in\mathcal{C}} h_{\mathsf{C}} S_{\mathsf{C}}| < \ell) \leq 2 \left[\sqrt{\frac{2}{\pi}} \frac{\ell}{\||\mathcal{C}|^{-1/2} S\|_2} + C_1 B \left(\frac{\||\mathcal{C}|^{-1/2} S\|_3}{\||\mathcal{C}|^{-1/2} S\|_2} \right)^3 \right]$$

where C_1 is an absolute constant, and $B = 1$ is the third moment of a Bernoulli random variable.

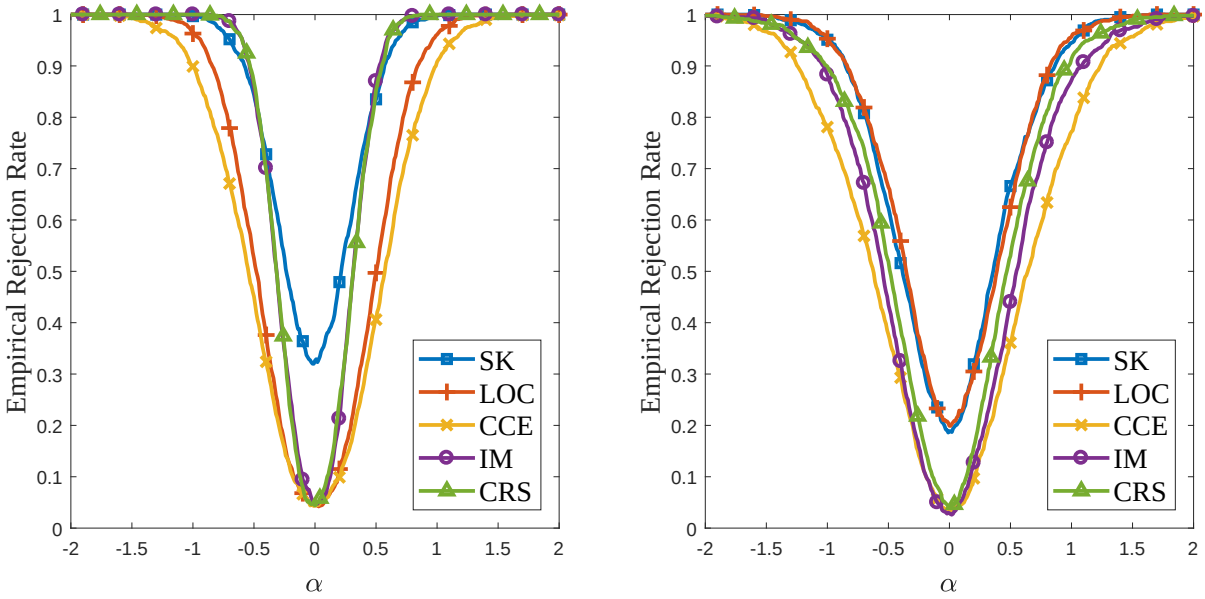
For a sequence $\check{G} \to \infty$, Gaussian concentration properties coupled with the assumed

bounds on the variances of the components of $\tilde{S}^*_{\mathcal{C}}$ imply that there is are sets $\mathcal{A}_{\mathcal{C}}$ closed under the action of $\mathcal{H}_{\mathcal{C}}$ with $\tilde{\Pr}(\mathcal{A}_{\mathcal{C}}) \to_{\check{G}\to\infty} 1$ and a fixed choice D which makes w Lipschitz with constant 1 on all of $\mathcal{A}_{\mathcal{C}}$, such that ℓ is bounded by a fixed constant times δ for all S in all $\mathcal{A}_{\mathcal{C}}$. Gaussian concentration properties for $\||\mathcal{C}|^{-1/2}\tilde{S}^*_{\mathcal{C}}\|_2$ and $(\||\mathcal{C}|^{-1/2}\tilde{S}^*_{\mathcal{C}}\|_3/\||\mathcal{C}|^{-1/2}\tilde{S}^*_{\mathcal{C}}\|_2)^3$ then also allow the construction of $\check{\delta}_n \to 0$ such that the right hand of the above expression side also $\to 0$ provided that $\check{G}$ grows sufficiently slowly.

B.5 Additional Simulation Results

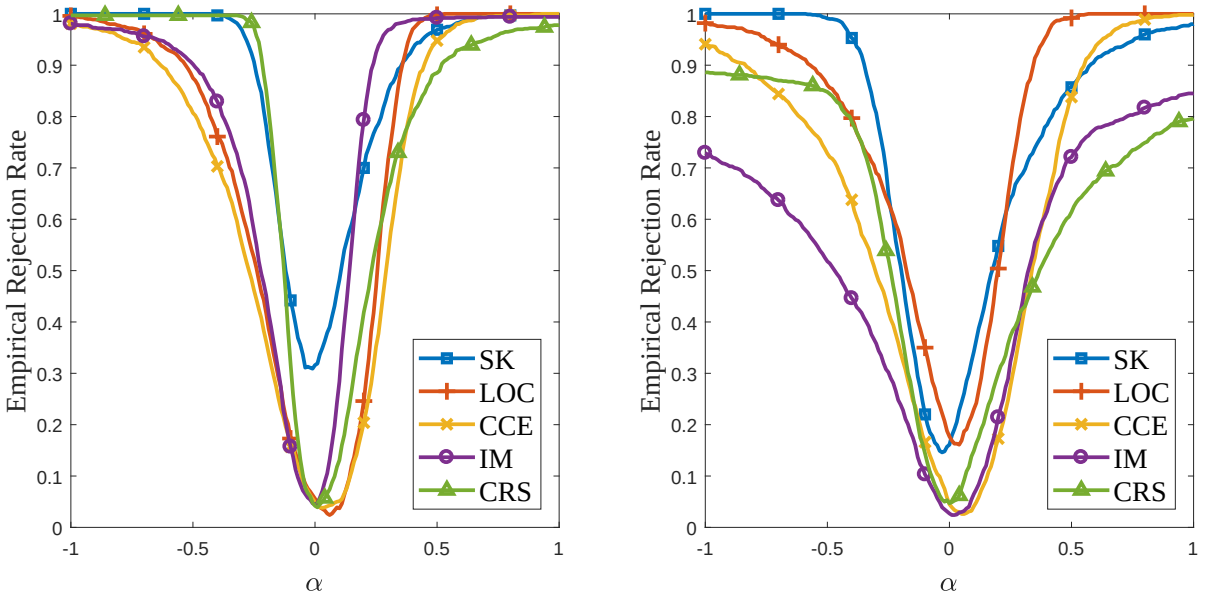
This section provides an additional set of simulation results. The same settings as in Section 2.4 are considered, except that the number of locations is $N = 820$ instead of $N = 205$. In those settings, four copies of the locations from the empirical example are created by reflecting the original locations over the 29° latitude and 75° longitude lines. The data generating process follows Section 2.4. The maximal number of groups to be considered in CCE, IM, and CRS is chosen to be $G_{\max} = 12 = \lceil (NT)^{1/3} \rceil$.

The results are shown in Table B.1, B.2, B.3, B.4 and Figure B.1, B.2 are qualitatively similar to the results from Section 2.4 (cases with $N = 205$). One major difference is that interior solutions for choosing $\widehat{G}$ are more likely to be chosen. For example, IM has 81.2% chance of choosing $G = 8$ as the optimal number of groups in the case of IV with SAR.



(a) BASELINE (b) SAR

Figure B.1: OLS power curves - $N = 820$



(a) BASELINE (b) SAR

Figure B.2: IV power curves - $N = 820$

Table B.1: Simulation Results: OLS - $N = 820$

Method	Estim. Bias	Estim. RMSE	Size	Power			
				-1	-0.5	0.5	1
A. BASELINE							
SK	0.007	0.253	0.326	0.997	0.844	0.835	0.998
LOC-U	0.007	0.253	0.682	1.000	0.959	0.939	1.000
LOC	0.007	0.253	0.046	0.963	0.525	0.497	0.961
CCE	0.007	0.253	0.044	0.899	0.451	0.406	0.908
IM	0.003	0.146	0.048	1.000	0.857	0.870	0.999
CRS	0.002	0.145	0.044	1.000	0.868	0.853	1.000
B. SAR							
SK	0.008	0.320	0.189	0.951	0.624	0.666	0.945
LOC-U	0.008	0.320	0.535	0.992	0.827	0.824	0.987
LOC	0.008	0.320	0.199	0.953	0.664	0.625	0.957
CCE	0.008	0.320	0.035	0.781	0.377	0.361	0.772
IM	0.001	0.275	0.029	0.883	0.435	0.440	0.872
CRS	0.005	0.256	0.040	0.899	0.519	0.530	0.914

Notes: Simulation results for estimation in the design described in Section 2.2. The nominal size is 0.05. Estimates are presented for the estimators, SK, LOC-U, LOC, CCE, IM, CRS described in the text. This table displays settings A and B described in the text. For each estimator and settings, columns display method, estimated bias, estimated RMSE, size, and power against 4 alternatives (-1, -0.5, 0.5, 1). Figures are based on 1000 simulation replications.

Table B.2: Simulation Results: IV - $N = 820$

Method	Estim. Median	Estim. MAD	Size	Power			
				-1	-0.5	0.5	1
A. BASELINE							
SK	0.003	0.085	0.316	1.000	1.000	0.969	1.000
LOC-U	0.003	0.085	0.682	1.000	0.991	1.000	1.000
LOC	0.003	0.085	0.056	0.996	0.875	0.998	1.000
CCE	0.003	0.085	0.049	0.981	0.807	0.948	1.000
IM	-0.049	0.068	0.044	0.981	0.902	0.992	0.994
CRS	-0.051	0.068	0.042	0.997	0.997	0.885	0.978
B. SAR							
SK	0.012	0.110	0.161	1.000	0.991	0.857	0.980
LOC-U	0.012	0.110	0.526	0.994	0.935	1.000	1.000
LOC	0.012	0.110	0.177	0.982	0.861	0.992	1.000
CCE	0.012	0.110	0.045	0.941	0.728	0.838	0.998
IM	-0.059	0.150	0.027	0.729	0.518	0.721	0.845
CRS	-0.042	0.141	0.047	0.886	0.848	0.614	0.796

Notes: Simulation results for estimation in the design described in Section 2.2. The nominal size is 0.05. Estimates are presented for the estimators, SK, LOC-U, LOC, CCE, IM, CRS described in the text. This table displays settings A and B described in the text. For each estimator and settings, columns display method, estimated median, estimated median absolute deviation, size, and power against 4 alternatives (-1, -0.5, 0.5, 1). Figures are based on 1000 simulation replications.

Table B.3: Clustering: OLS - $N = 820$

		G											$\widehat{G}$
		2	3	4	5	6	7	8	9	10	11	12	
A. BASELINE													
CCE	size	0.041	0.052	0.043	0.048	0.044	0.047	0.037	0.035	0.048	0.043	0.047	0.044
	$\widehat{G}$ frequency	0.000	0.002	0.015	0.046	0.100	0.122	0.107	0.098	0.141	0.134	0.235	-
IM	size	0.047	0.042	0.039	0.042	0.041	0.041	0.048	0.036	0.053	0.038	0.037	0.048
	$\widehat{G}$ frequency	0.000	0.000	0.000	0.000	0.000	0.000	0.014	0.008	0.179	0.162	0.637	-
CRS	size	0.000	0.000	0.000	0.000	0.025	0.040	0.041	0.033	0.051	0.040	0.039	0.044
	$\widehat{G}$ frequency	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.009	0.153	0.162	0.672	-
B. SAR													
CCE	size	0.064	0.041	0.040	0.042	0.031	0.032	0.025	0.024	0.036	0.035	0.039	0.035
	$\widehat{G}$ frequency	0.000	0.000	0.000	0.008	0.043	0.069	0.091	0.121	0.129	0.199	0.340	-
IM	size	0.050	0.032	0.048	0.038	0.026	0.027	0.025	0.030	0.036	0.042	0.042	0.029
	$\widehat{G}$ frequency	0.000	0.000	0.000	0.011	0.007	0.151	0.454	0.176	0.013	0.014	0.174	-
CRS	size	0.000	0.000	0.000	0.000	0.036	0.042	0.045	0.045	0.053	0.050	0.049	0.040
	$\widehat{G}$ frequency	0.000	0.000	0.000	0.000	0.000	0.095	0.246	0.163	0.039	0.054	0.403	-

Notes: Simulation results for the design described in Section 2.2. Inferential properties are presented are presented for the estimators, CCE, IM, CRS, described in the text. This table displays settings A and B described in the text. $\widehat{G}$ is the number of clusters chosen by the criterion on size-power tradeoff described in the text. The rows "$\widehat{G}$ frequency" is the frequency of a particular G achieving the highest simulated power among candidate G's in the setting.

Table B.4: Clustering: IV - $N = 820$

		G											$\widehat{G}$
		2	3	4	5	6	7	8	9	10	11	12	
A. BASELINE													
CCE	size	0.049	0.053	0.043	0.059	0.051	0.050	0.044	0.044	0.051	0.043	0.046	0.049
	$\widehat{G}$ frequency	0.000	0.001	0.003	0.035	0.042	0.058	0.073	0.109	0.128	0.140	0.411	-
IM	size	0.050	0.042	0.040	0.046	0.038	0.041	0.046	0.034	0.046	0.036	0.041	0.044
	$\widehat{G}$ frequency	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.000	0.179	0.067	0.748	-
CRS	size	0.000	0.000	0.000	0.000	0.025	0.036	0.041	0.027	0.050	0.038	0.040	0.042
	$\widehat{G}$ frequency	0.000	0.000	0.000	0.000	0.000	0.000	0.008	0.007	0.199	0.133	0.653	-
B. SAR													
CCE	size	0.049	0.049	0.037	0.045	0.039	0.036	0.034	0.036	0.041	0.042	0.046	0.045
	$\widehat{G}$ frequency	0.000	0.001	0.000	0.002	0.009	0.030	0.045	0.094	0.134	0.200	0.485	-
IM	size	0.045	0.029	0.045	0.034	0.019	0.021	0.024	0.029	0.021	0.026	0.024	0.027
	$\widehat{G}$ frequency	0.000	0.000	0.013	0.081	0.008	0.074	0.812	0.011	0.000	0.000	0.001	-
CRS	size	0.000	0.000	0.000	0.000	0.037	0.033	0.043	0.048	0.039	0.039	0.041	0.047
	$\widehat{G}$ frequency	0.000	0.000	0.000	0.000	0.000	0.095	0.568	0.265	0.020	0.010	0.042	-

Notes: Simulation results for the design described in Section 2.2. Inferential properties are presented are presented for the estimators, CCE, IM, CRS, described in the text. This table displays settings A and B described in the text. $\widehat{G}$ is the number of clusters chosen by the criterion on size-power tradeoff described in the text. The rows "$\widehat{G}$ frequency" is the frequency of a particular G achieving the highest simulated power among candidate G's in the setting.

APPENDIX C

APPENDIX FOR CHAPTER 3

Proof of Lemma 3.1. (i) First assume Condition IN and Condition ST holds. The proof follows Ferman and Pinto (2019b), except that we do not assume that there is a set of weights that reconstruct the factor loadings and belong to the simplex.

We first show part (b). It suffices to show $|\widehat{a}_i - a_i| = o_p(1)$ and $\|\widehat{b}_i - b_i\| = o_p(1)$ for each i, i.e. a_i and b_i are well-defined. We show it for the $i = 1$ case and other cases follow the same strategy. Let $\bar{y}_j = T^{-1}\sum_{t=1}^{T} y_{j,t}$. Write down an (equivalent) optimization problem

$$\widehat{v} = \underset{v \in V}{\arg\min} \left((y_{1,t} - \bar{y}_1) - \sum_{j=2}^{N} (y_{j,t} - \bar{y}_j) v_j \right)^2,$$

where $V = \{v = (v_2, \ldots, v_N) \in \mathbb{R}_+^{N-1} : \sum_{j=2}^{N} v_j = 1\}$. The objective is strictly convex (with probability approaching one), so the solution is unique. Note that it implies $\widehat{b}_1$ is numerically equivalent to $(0, \widehat{v}')'$, otherwise the minimization problem in forming $\widehat{a}_1$ and $\widehat{b}_1$ may have a lower objective evaluated at $(\bar{y}_1 - \sum_{j=2}^{N} \bar{y}_j \widehat{v}_j, 0, \widehat{v}')'$. Now we let $\widehat{Q}(v)$ denote the objective function such that

$$\widehat{Q}(v) = \frac{1}{T}\sum_{t=1}^{T} \left((y_{1,t} - \bar{y}_1) - \sum_{j=2}^{N} (y_{j,t} - \bar{y}_j) v_j \right)^2,$$

and its population analog be

$$Q(v) = \begin{bmatrix} -1 \\ v \end{bmatrix}' \Omega_y \begin{bmatrix} -1 \\ v \end{bmatrix}.$$

Let v_0 be a minimizer of $Q(v)$ in V. We verify the conditions for consistency (see Newey and McFadden, 1994, Theorem 2.1) : (i) Since Ω_y is positive definite, $Q(v)$ is strictly convex.

Also, V is convex. Therefore, $Q(v)$ is uniquely minimized at v_0. (ii) V is compact, since it is a $(N-1)$-dimensional simplex. (iii) $Q(v)$ is continuous, since it has a quadratic form. (iv) To see uniform convergence, note

$$\begin{aligned}
\sup_{v \in V} |\widehat{Q}(v) - Q(v)| &= \sup_{v \in V} \left| \begin{bmatrix} -1 \\ v \end{bmatrix}' \left(\frac{1}{T} \sum_{t=1}^{T} (Y_t - \bar{Y})(Y_t - \bar{Y})' - \Omega_y \right) \begin{bmatrix} -1 \\ v \end{bmatrix} \right| \\
&\leq \sup_{v \in V} \left\| \begin{bmatrix} -1 \\ v \end{bmatrix} \right\|^2 \left\| \frac{1}{T} \sum_{t=1}^{T} (Y_t - \bar{Y})(Y_t - \bar{Y})' - \Omega_y \right\|_F \\
&\leq N \cdot o_p(1) \\
&= o_p(1),
\end{aligned}$$

where $\| \cdot \|_F$ is the Frobenius norm. The second inequality is by ergodicity for the second moments. Therefore, $\widehat{v} \to_p v_0$. This implies $\|\widehat{b}_1 - b_1\| = o_p(1)$. By ergodicity,

$$\widehat{a}_1 = \bar{y}_1 - [\bar{y}_2 \ \bar{y}_3 \ \dots \ \bar{y}_N]\widehat{v} \to_p E[y_{1,t}(0) - Y_t(0)' b_1] = a_1.$$

This shows part (b) and $E[u_{1,t}] = 0$ by definition of $u_{i,t}$. We also have that $\{u_t\}_{t \geq 1}$ is stationary since it is a linear combination of stationary and ergodic processes. This shows part (a) in Assumption 3.1.

Part (c) follows from part (b) and the stationarity of $\{Y_{T+1}(0)\}_{T \geq 1}$. Part (d) follows by Condition IN. Thus, Assumption 3.1 holds under Condition IN and Condition ST.

(ii) Now we instead assume Condition IN and Condition CO holds.

We first show part (c). We will show $\|Y_{T+1}(0)'(\widehat{b}_1 - b_1)\| = o_p(1)$ and other i's follows the same strategy. Since the synthetic control estimator can be written as a projection of the OLS estimator onto a closed convex set, we will first derive the asymptotic properties of

the OLS estimator, and then use the properties of projections to obtain the desired results. For examples of this strategy, see Li (2019) and Yu et al. (2019). For some positive definite matrix $D \in \mathbb{R}^N$, let $\mathbb{R}^N$ be a Hilbert space with the inner product $\langle \cdot, \cdot \rangle_D$ such that for $\theta_1, \theta_2 \in \mathbb{R}^N$,

$$\langle \theta_1, \theta_2 \rangle_D = \theta_1' D \theta_2.$$

The norm $\| \cdot \|_D$ is defined accordingly, i.e. $\|\theta\|_D = \sqrt{\theta' D \theta}$, for $\theta \in \mathbb{R}^N$. For a closed convex set $\Lambda \subset \mathbb{R}^N$, define a projection Π_D such that for each $\theta \in \mathbb{R}^N$, $\Pi_D \theta = \arg\min_{\theta' \in \Lambda} \|\theta - \theta'\|_D$. Zarantonello (1971) shows that for each $\theta, \theta' \in \mathbb{R}^N$,

$$\|\Pi_D \theta - \Pi_D \theta'\|_D \leq \|\theta - \theta'\|_D. \tag{C.1}$$

With some abuse of notation, let $x_t = Y_t - T^{-1} \sum_{s=1}^T Y_s$. Then, $\widehat{b}_1$ is the synthetic control weight estimators of regressing $(y_{1,t} - T^{-1} \sum_{s=1}^T y_{1,s})$ on x_t, subject to $\{0\} \times \Delta_{N-1}$ with Δ_{N-1} being an $(N-1)$-dimensional simplex. Let $\tilde{b}_1$ be the OLS estimator of regressing $(y_{1,t} - T^{-1} \sum_{s=1}^T y_{1,s})$ on x_t. Let $\Sigma_T = T^{-1} \sum_{t=1}^T x_t x_t'$.

Appendix A.2 in Li (2019) establishes that $\widehat{b}_1 = \Pi_{\Sigma_T} \tilde{b}_1$. Thus, we have

$$\begin{aligned}
\|\widehat{b}_1 - b_1\| &= \|\Sigma_T^{-1/2} \Sigma_T^{1/2} (\widehat{b}_1 - b_1)\| \\
&\leq \|\Sigma_T^{-1/2}\|_F \cdot \|\Sigma_T^{1/2} (\widehat{b}_1 - b_1)\| \\
&= \|\Sigma_T^{-1/2}\|_F \cdot \|\widehat{b}_1 - b_1\|_{\Sigma_T} \\
&= \|\Sigma_T^{-1/2}\|_F \cdot \|\Pi_{\Sigma_T} \tilde{b}_1 - \Pi_{\Sigma_T} b_1\|_{\Sigma_T} \\
&\leq \|\Sigma_T^{-1/2}\|_F \cdot \|\tilde{b}_1 - b_1\|_{\Sigma_T} \\
&= \|\Sigma_T^{-1/2}\|_F \cdot \|\Sigma_T^{1/2}\|_F \cdot \|\tilde{b}_1 - b_1\| \\
&= O_p(1) o_p(T^{-1/2}) \\
&= o_p(T^{-1/2}),
\end{aligned} \tag{C.2}$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix. The third equality is because $b_1 \in \{0\}\times\Delta_{N-1}$. The second inequality is by (C.1). To see the fifth equality, note

$$\Sigma_T = T\left(\frac{1}{T^2}\sum_{t=1}^{T} Y_t Y_t' - \left(\frac{1}{T^{3/2}}\sum_{t=1} Y_t\right)\left(\frac{1}{T^{3/2}}\sum_{t=1} Y_t\right)'\right),$$

so

$$\|\Sigma_T^{-1/2}\|_F \cdot \|\Sigma_T^{1/2}\|_F = \mathrm{tr}(\Sigma_T^{-1})\mathrm{tr}(\Sigma_T) = O_p(1)\cdot\frac{1}{T}\cdot T\cdot O_p(1) = O_p(1),$$

where the second equality is standard results for $\mathcal{I}_1$ process (see Hamilton, 1994, part (g) and (i) of Proposition 18.1). Also, $\|\tilde{b}_1 - b_1\| = o_p(T^{-1/2})$ is by Proposition 19.2 in Hamilton (1994). This shows (C.2). Apply part (a) of Proposition 18.1 in Hamilton (1994), we have

$$\|Y_{T+1}(0)'(\widehat{b}_1 - b)\| = \|(T^{-1/2}Y_{T+1}(0))'(T^{-1/2}(\widehat{b}_1 - b))\| = O_p(1)o_p(1) = o_p(1).$$

Now we show part (b). Again, it suffices to show $|\widehat{a}_i - a_i| = o_p(1)$ and $\|\widehat{b}_i - b_i\| = o_p(1)$. We consider the $i = 1$ case and other cases follow the same strategy. We have showed $\|\widehat{b}_i - b_i\| = o_p(1)$ in part (c) of the proof. Section A.6.1 in Ferman and Pinto (2019b) establishes that

$$[\mu_1^1\ \mu_2^1\ \ldots\ \mu_N^1](b_1 - e_1) = 0, \tag{C.3}$$

where e_i is the unit vector with one at the i-th entry. Thus,

$$\begin{aligned}
\widehat{a}_1 &= [\bar{y}_1\ \bar{y}_2\ \ldots\ \bar{y}_N](e_1 - \widehat{b}_1) \\
&= [\bar{y}_1\ \bar{y}_2\ \ldots\ \bar{y}_N](e_1 - b_1) + [\bar{y}_1\ \bar{y}_2\ \ldots\ \bar{y}_N](b_1 - \widehat{b}_1) \\
&= \left\{\frac{1}{T}\sum_{t=1}^{T}\left((\lambda_t^0)'[\mu_1^0\ \ldots\ \mu_N^0] + [\epsilon_{1,t}\ \ldots\ \epsilon_{N,t}]\right)\right\}(e_1 - b_1) + \\
&\quad \left(\frac{1}{\sqrt{T}}[\bar{y}_1\ \bar{y}_2\ \ldots\ \bar{y}_N]\right)\sqrt{T}(b_1 - \widehat{b}_1)
\end{aligned}$$

$$
\begin{aligned}
&= E[\lambda_t^0]'[\mu_1^0 \ \dots \ \mu_N^0](e_1 - b_1) + o_p(1) + O_p(1)o_p(1) \\
&\to_p E[\lambda_t^0]'[\mu_1^0 \ \dots \ \mu_N^0](e_1 - b_1). \\
&= a_1 \qquad \text{(C.4)}
\end{aligned}
$$

The third equality is by (C.3). The fourth equality is by stationarity of $\{(\lambda_t^0, \epsilon_t)\}_{t\geq 1}$ and results in part (d) of the proof. This shows part (b) of the Assumption 3.1 .

Combining (C.3) and (C.4), we have part (a) in Assumption 3.1. Part (d) is assumed by Condition IN. □

Proof of Theorem 3.1. Using formula of $\widehat{\gamma}$ in Equation (3.6), we have

$$
\begin{aligned}
\widehat{\gamma} &= (A'\widehat{M}A)^{-1}A'(I - \widehat{B})'((I - \widehat{B})Y_{T+1}(0) + (I - \widehat{B})\alpha - \widehat{a}) \\
&= (A'\widehat{M}A)^{-1}A'(I - \widehat{B})'(u_{T+1} + (B - \widehat{B})Y_{T+1}(0) + (a - \widehat{a}) + (I - \widehat{B})A\gamma) \\
&= (A'\widehat{M}A)^{-1}A'(I - \widehat{B})'u_{T+1} + o_p(1) + o_p(1) + \gamma.
\end{aligned}
$$

The first equality is by $Y_{T+1} = Y_{T+1}(0) + \alpha$. The second equation is because $Y_{T+1}(0) = a + BY_{T+1}(0) + u_{T+1}$. The third equation is by (b) and (c) in Assumption 3.1. Therefore,

$$
\begin{aligned}
\widehat{\alpha} - (\alpha + Gu_{T+1}) &= A(A'\widehat{M}A)^{-1}A'(I - \widehat{B})'u_{T+1} + A\gamma + o_p(1) - \alpha - Gu_{T+1} \\
&= (A(A'\widehat{M}A)^{-1}A'(I - \widehat{B}) - G)'u_{T+1} + o_p(1) \\
&= o_p(1)O_p(1) + o_p(1) \\
&= o_p(1).
\end{aligned}
$$

The third equality is by (b) in Assumption 3.1 and stationarity of $\{u_t\}_{t\geq 1}$. □

Proof of Proposition 3.1. The proof for the first half of the proposition is similar to the

proof for Theorem 3.1, and thus is omitted. Too see the second half, note

$$Cov[G_W u_{T+1}] = A(Q'WQ)^{-1}Q'W\Omega WQ(Q'WQ)^{-1}A'$$

and

$$Cov[G^e u_{T+1}] = A(Q'\Omega Q)^{-1}A',$$

where $Q = (I-B)A$. It suffices to show that $((Q'WQ)^{-1}Q'W\Omega WQ(Q'WQ)^{-1} - (Q'\Omega Q)^{-1})$ is positive semi-definite. Note that the first term is asymptotic variance of using W as the weighting matrix in GMM exercise and the second term is the one using the efficient weighting matrix (see Hayashi, 2000, Proposition 3.5). Thus, $(Cov[G_W u_{T+1}] - Cov[G^e u_{T+1}])$ is positive semi-definite. □

Proof of Lemma 3.2. Since Assumption 3.3 implies Assumption 3.2, we only need to show Lemma 3.3. □

Proof of Theorem 3.2. We follow the proof of Theorem 2 in Andrews et al. (2006). Let

$$L_{1,T}(\epsilon) = \left\{ \|C_T(\widehat{\beta}_1 - \beta_1)\| \leq \epsilon, \max_{t=1,\ldots,T} \|C_T(\widehat{\beta}_1^{(t)} - \beta_1)\| \leq \epsilon \right\},$$
$$L_{2,T}(c) = \left\{ \max_{t \leq T+1} \|C_T^{-1} x_t\| \leq c \right\}.$$

By Assumption 3.2(d), there exists a positive sequence $\{\epsilon_T\}_{T\geq 1}$ such that $\epsilon_T \to 0$ and $\Pr(L_{1,T}(\epsilon_T)) \to 1$. Let $c_T = 1/\sqrt{\epsilon_T}$. So we have $c_T \to \infty$ and $c_T\epsilon_T \to 0$. By Assumption 3.2(c), we must have $\Pr(L_{2,T}(c_T)) \to 1$. Let $L_T = L_{1,T}(\epsilon_T) \cap L_{2,T}(c_T)$, then we have $\Pr(L_T) \to 1$ and $\Pr(L_T^c) \to 0$.

Suppose L_T holds. Then, for $\beta = \widehat{\beta}_1$ or $\beta = \widehat{\beta}_1^{(t)}$ for some $t = 1, \ldots, T$, we have

$$\begin{aligned} |P_t(\beta) - P_t(\beta_1)| &= \left|(\beta-\beta_1)' x_t x_t' (\beta-\beta_1) - 2x_t'(\beta-\beta_1)u_{1,t}\right| \\ &= \left|(\beta-\beta_1)' C_T' (C_T')^{-1} x_t x_t' C_T^{-1} C_T (\beta-\beta_1) - 2x_t' C_T^{-1} C_T (\beta-\beta_1) u_{1,t}\right| \end{aligned}$$

$$
\begin{aligned}
&\leq \|C_T(\beta-\beta_1)\|^2\|C_T^{-1}x_t\|^2 + 2\|C_T^{-1}x_t\|\|C_T(\beta-\beta_1)\||u_{1,t}| \\
&\leq \epsilon_T^2 c_T^2 + 2\epsilon_T c_T |u_{1,t}|.
\end{aligned}
$$

Define $g_t(\epsilon_T, c_T) = \epsilon_T^2 c_T^2 + 2\epsilon_T c_T|u_{1,t}|$. Note that $g_t(\epsilon_T, c_T)$ is identically distributed across t for a fixed T, by Assumption 3.2(a).

We first prove part (a). Let x be some continuous point of distribution function of $P_{T+1}(\beta_1)$. Then,

$$
\begin{aligned}
\Pr(P_{T+1}(\widehat{\beta}_1) \leq x) &= \Pr(\{P_{T+1}(\widehat{\beta}_1) \leq x\} \cap L_T) + \Pr(\{P_{T+1}(\widehat{\beta}_1) \leq x\} \cap L_T^c) \\
&\leq \Pr(P_{T+1}(\widehat{\beta}_1) \leq x + g_t(\epsilon_T, c_T)) + \Pr(L_T^c) \\
&\leq \Pr(P_{T+1}(\beta_1) \leq x) + o(1).
\end{aligned}
$$

To see the last equality, pick $\epsilon > 0$. By continuity, $\exists \delta > 0$ such that for each $y \in (x-\delta, x+\delta)$, $|\Pr(P_{T+1}(\beta_1) \leq y) - \Pr(P_{T+1}(\beta_1) \leq x)| < \epsilon$. Therefore,

$$
\begin{aligned}
\Pr(P_{T+1}(\widehat{\beta}_1) \leq x + g_t(\epsilon_T, c_T)) &= \Pr(\{P_{T+1}(\widehat{\beta}_1) \leq x + g_t(\epsilon_T, c_T)\} \cap \{|g_t(\epsilon_T, c_T)| \geq \delta\}) \\
&\quad + \Pr(\{P_{T+1}(\widehat{\beta}_1) \leq x + g_t(\epsilon_T, c_T)\} \cap \{|g_t(\epsilon_T, c_T)| < \delta\}) \\
&\leq \Pr(|g_t(\epsilon_T, c_T)| \geq \delta) + \Pr(P_{T+1}(\widehat{\beta}_1) \leq y) \\
&< \Pr(P_{T+1}(\beta_1) \leq x) + o(1).
\end{aligned}
$$

Similarly,

$$
\Pr(P_{T+1}(\widehat{\beta}_1) \leq x) \geq \Pr(P_{T+1}(\beta_1) \leq x) + o(1).
$$

This shows part (a).

To see part (b), let $k : \mathbb{R} \to \mathbb{R}$ be a monotonically decreasing and everywhere differentiable function that has bounded derivative and satisfies $k(x) = 1$ for $x \leq 0$, $k(x) \in [0,1]$ for $x \in (0,1)$, and $k(x) = 0$ for $x \geq 1$. For example, let $k(x) = \cos(\pi x)/2 + 1/2$ for $x \in (0,1)$.

Given some $\{\beta^{(t)}\}_{t=1}^{T}$, a smoothed df is defined by

$$\widehat{F}_T(x, \{\beta^t\}, h_T) = \frac{1}{T}\sum_{t=1}^{T} k\left(\frac{P_t(\beta^{(t)}) - x}{h_T}\right),$$

for some sequence of positive constants $\{h_T\}$ such that $h_T \to 0$ and $c_T\epsilon_T/h_T \to 0$. For example, we let $h_T = \epsilon_T^{1/4}$ when $c_T = 1/\sqrt{\epsilon_T}$. Also, define,

$$\widehat{F}_T(x, \{\beta_1\}) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{P_t(\beta_1) \le x\},$$

i.e., $\widehat{F}_T(x, \{\beta_1\})$ is the empirical cdf of P_t as if the true parameter β_1 is known.

We write

$$|\widehat{F}_{P,T}(x) - F_P(x)| \le \sum_{i=1}^{4} D_{i,T},$$

for

$$\begin{aligned}
D_{1,T} &= |\widehat{F}_{P,T}(x) - \widehat{F}_T(x, \{\widehat{\beta}_j\}, h_T)|,\\
D_{2,T} &= |\widehat{F}_T(x, \{\widehat{\beta}_j\}, h_T) - \widehat{F}_T(x, \{\beta_1\}, h_T)|,\\
D_{3,T} &= |\widehat{F}_T(x, \{\beta_1\}, h_T) - \widehat{F}_T(x, \{\beta_1\})|, \text{ and}\\
D_{4,T} &= |\widehat{F}_T(x, \{\beta_1\}) - F_P(x)|.
\end{aligned}$$

We want to show that all four terms vanish. First note that

$$D_{1,T} \le \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left\{\frac{P_t(\widehat{\beta}_1^{(t)}) - x}{h_T} \in (0,1)\right\}.$$

Thus, for any $\delta > 0$,

$$\Pr(D_{1,T} > \delta) \le \Pr(\{D_{1,T} > \delta\} \cap L_T) + \Pr(L_T^c)$$

$$\leq \Pr\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}\left\{P_t(\widehat{\beta}_1^{(t)}) - x \in (-g_t(\epsilon_T, c_T), h_T + g_t(\epsilon_T, c_T)\right\} > \delta\right) + o(1)$$

$$\leq \frac{E\mathbb{1}\left\{P_t(\widehat{\beta}_1^{(t)}) - x \in (-g_t(\epsilon_T, c_T), h_T + g_t(\epsilon_T, c_T)\right\}}{\delta} + o(1), \tag{C.5}$$

where the last inequality is by Markov's inequality. Recall $\Pr(P_1(\beta_1) \neq x) = 1$ and $g_t(\epsilon_T, c_T) \to 0$ a.s., so $\mathbb{1}\{P_t(\beta_1) - x \in \{-g_t(\epsilon_T, c_T), h_T + g_t(\epsilon_T, c_T)\} \to 0$ a.s.. By the dominated convergence theorem, (C.5) implies $\Pr(D_{1,T} > \delta) \leq o(1)$ and thus $D_{1,T} = o_p(1)$.

For $D_{2,T}$, we have

$$\begin{aligned} D_{2,T} &= \left|\frac{1}{T}\sum_{t=1}^{T} k'\left(\frac{\tilde{P}_t - x}{h_T}\right)\frac{P_t(\widehat{\beta}_1^{(t)}) - P_t(\beta_1)}{h_T}\right| \\ &\leq \frac{\bar{k}}{T}\sum_{t=1}^{T}\frac{g_t(\epsilon_T, c_T)}{h_T}. \end{aligned}$$

The equality is by the mean value theorem and we have $\tilde{P}_t$ lies between $P_t(\widehat{\beta}_1^{(t)})$ and $P_t(\beta_1)$. In the inequality, $\bar{k}$ is a bound for the derivative of k. Also, note

$$E\left[\frac{g_t(\epsilon_T, c_T)}{h_T}\right] = \frac{\epsilon_T^2 c_T^2}{h_T} + 2\frac{\epsilon_T c_T}{h_T}E|u_{1,t}| = o(1).$$

Therefore,

$$\begin{aligned} \Pr(D_{2,T} > \delta) &\leq \Pr(\{D_{2,T} > \delta\} \cap L_T) + \Pr(L_T^c) \\ &\leq \Pr\left(\frac{\bar{k}}{T}\sum_{t=1}^{T}\frac{g_t(\epsilon_T, c_T)}{h_T} > \delta\right) + o(1) \\ &\leq \bar{k}\frac{Eg_t(\epsilon_T, c_T)}{\delta h_T} \\ &\to 0. \end{aligned}$$

The third inequality is by Markov's inequality. This shows $D_{2,T} = o_p(1)$.

$D_{3,T}$ is similar to the $D_{1,T}$ case. Finally, by stationary and ergodicity of $u_{1,t}$, we have $D_{4,T} = o_p(1)$. This shows part (b).

Now we show part (c). Pick any small ϵ such that $\widehat{F}_{P,T}(x) \to_p F_P(x)$ for $x \in (q_{P,1-\tau} - \epsilon, q_{P,1-\tau} + \epsilon)$. Note

$$\begin{aligned}\Pr(\widehat{q}_{P,1-\tau} > q_{P,1-\tau} + \epsilon) &\leq \Pr(\widehat{F}_{P,T}(q_{P,1-\tau} + \epsilon) < 1 - \tau) \\ &= \Pr(\widehat{F}_{P,T}(q_{P,1-\tau} + \epsilon) - F_P(q_{P,1-\tau} + \epsilon) < (1 - \tau) - F_P(q_{P,1-\tau} + \epsilon)) \\ &\to 0.\end{aligned}$$

The inequality is by definition of $\widehat{q}_{P,1-\tau}$. The convergence is because of part (e) of Assumption 3.2 and part (b) of Theorem 3.2. Similarly,

$$\begin{aligned}&\Pr(\widehat{q}_{P,1-\tau} < q_{P,1-\tau} - \epsilon) \\ &\leq \Pr(\widehat{F}_{P,T}(q_{P,1-\tau} - \epsilon) \geq 1 - \tau) \\ &= \Pr(\widehat{F}_{P,T}(q_{P,1-\tau} - \epsilon) - F_P(q_{P,1-\tau} - \epsilon) \geq (1 - \tau) - F_P(q_{P,1-\tau} - \epsilon)) \\ &\to 0.\end{aligned}$$

Again, the inequality is by definition of $\widehat{q}_{P,1-\tau}$, and the convergence is because of part (e) of Assumption 3.2 and part (b) of Theorem 3.2.

Finally, we show part (d). Under null, P_∞ and $P_1(\beta_1)$ have the same distribution, so $q_{P,1-\tau}$ is $(1-\tau)$-quantile of P_∞. Therefore,

$$\begin{aligned}\Pr(P > \widehat{q}_{P,1-\tau}) &= 1 - \Pr(P \leq \widehat{q}_{P,1-\tau}) \\ &= 1 - \Pr(P + (q_{P,1-\tau} - \widehat{q}_{P,1-\tau}) \leq q_{P,1-\tau}) \\ &\to \tau,\end{aligned}$$

where the convergence is by combining part (a) and (c). This concludes our proof. □

Proof of Lemma 3.3. (i) Assume Condition ST holds.

By Lemma 3.1, part (a) of Assumption 3.3 holds.

Part (b) is because u_t is a linear combination of $\eta_t, \lambda_t, \epsilon_t$.

For part (c), pick some τ such that $1/(2+\delta) < \tau < 1/2$, where δ is defined in Condition ST. Let

$$D_T = \begin{bmatrix} 1 & 0 \\ 0 & T^{\tau} I_N \end{bmatrix}. \tag{C.6}$$

Then, we have

$$\max_{t \leq T+1} \|D_T^{-1} x_t\| = \max_{t \leq T+1} \left\| \begin{bmatrix} 1 \\ T^{-\tau} Y_t \end{bmatrix} \right\| = \sqrt{1 + \left(\max_{t \leq T+1} \|T^{-\tau} Y_t\| \right)^2}. \tag{C.7}$$

Also, for any $\epsilon > 0$, note

$$\begin{aligned} \Pr\left(\max_{t \leq T+1} \|T^{-\tau} Y_t\| > \epsilon \right) &= \Pr\left(\bigcup_{t \leq T+1} \|Y_t\| > T^{\tau} \epsilon \right) \\ &\leq \left(\sum_{t=1}^{T} \Pr(\|Y_t\| > T^{\tau} \epsilon) \right) + \Pr(\|Y_{T+1}(0) + \alpha\| > T^{\tau} \epsilon) \\ &= \frac{T E[\|Y_t\|^{2+\delta}]}{T^{\tau(2+\delta)} \epsilon^{2+\delta}} + o(1) \\ &= o(1). \end{aligned} \tag{C.8}$$

The second equality is due to Markov inequality and stationarity of $\{Y_{T+1}(0)\}_{t+1}$. The last equality is because $\tau > 1/(2+\delta)$. Combining (C.7) and (C.8), we obtain part (c).

For part (d), we use D_T defined in (C.6). Following the same reasoning as in (C.2), for

each $i = 1, \ldots, N$, we have

$$\begin{aligned}\|\widehat{b}_i - b_i\| &\leq \|\Sigma_T^{-1/2}\|_F \cdot \|\Sigma_T^{1/2}\|_F \cdot \|\tilde{b}_i - b_i\| \\ &= O_p(1)O_p(T^{-1/2}) \\ &= O_p(T^{-1/2}). \end{aligned} \tag{C.9}$$

The first equality is because $\{Y_t(0)\}_{t\geq 1}$ is ergodic for the second moment, and $\tilde{b}_i$ is the OLS estimator for b_i. Thus,

$$\begin{aligned}\|D_T(\widehat{\beta}_i - \beta_i)\| &= \left\| \begin{bmatrix} 1 & 0 \\ 0 & T^{\tau-1/2}I_N \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & T^{1/2}I_N \end{bmatrix} (\widehat{\beta}_i - \beta_i) \right\| \\ &\leq \left\| \begin{bmatrix} 1 & 0 \\ 0 & T^{\tau-1/2}I_N \end{bmatrix} \right\|_F \left\| \begin{bmatrix} \widehat{a}_i - a_i \\ \sqrt{T}(\widehat{b}_i - b_i) \end{bmatrix} \right\| \\ &= \sqrt{1 + NT^{2\tau-1}} \|O_p(1)\| \\ &= o_p(1). \end{aligned}$$

The second equality is due to (C.9). The last equality is because $\tau < 1/2$. Therefore,

$$\|(\widehat{\theta} - \theta_0)D_T\|_F = \sqrt{\sum_{i=1}^{N} \|D_T(\widehat{\beta}_i - \beta_i)\|^2} = o_p(1).$$

Also, since $\widehat{\theta}^{(t)} = \widehat{\theta}$ for each t,

$$\max_{t=1,\ldots,T} \|(\widehat{\theta}^{(t)} - \theta_0)D_T\|_F = \|(\widehat{\theta} - \theta_0)D_T\|_F = o_p(1).$$

This shows part (d).

Part (e) is assumed.

Part (f) is trivial if $W_T = I$. Assume now $W_T = (C\widehat{G}(T^{-1}\sum_{t=1}^T \widehat{u}_t\widehat{u}_t')\widehat{G}'C')^{-1}$. Then,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^T \widehat{u}_t\widehat{u}_t' &= (I-\widehat{B})\left(\frac{1}{T}\sum_{t=1}^T Y_tY_t'\right)(I-\widehat{B})' - (I-\widehat{B})\left(\frac{1}{T}\sum_{t=1}^T Y_t\right)\widehat{a}' - \\ &\quad \widehat{a}\left(\frac{1}{T}\sum_{t=1}^T Y_t'\right)(I-\widehat{B})' + \widehat{a}\widehat{a}' \\ &\to E[u_tu_t'],\end{aligned}$$

by ergodicity and Assumption 3.1(b). Therefore, $\widehat{W}_T \to_p W = (CGE[u_tu_t']G'C')^{-1}$.

This concludes part (i) of Lemma 3.3.

(ii) Assume Condition CO holds.

By Lemma 3.1, Assumption 3.1 holds. This shows Part (a).

By (C.3), u_t is a linear combination of λ_t^o and ϵ_t, so $\{u_t\}_{t\geq 1}$ is ergodic and has finite first moment. This shows Part (b).

Now we show Part (c). Let

$$D_T = \begin{bmatrix} 1 & 0 \\ 0 & \sqrt{T}\cdot I_N \end{bmatrix}.$$

Then, we have

$$\begin{aligned}\max_{t\leq T+1}\|D_T^{-1}x_t\| &= \sqrt{1+\left(\max_{t\leq T+1}\|T^{-1/2}Y_t\|\right)^2} \\ &\leq \sqrt{1+\sum_{i=1}^N\left(\max_{t\leq T+1}|T^{-1/2}y_{i,t}|\right)^2} \\ &\leq \sqrt{1+\sum_{i=1}^N\left(T^{-1/2}|\alpha_i| + \max_{t\leq T+1}|T^{-1/2}y_{i,t}(0)|\right)^2}\end{aligned}$$

$$= \sqrt{1 + \sum_{i=1}^{N} \big(o(1) + O_p(1)\big)^2}$$
$$= O_p(1)$$

The second equality is because

$$\max_{t \le T+1} |T^{-1/2} y_{i,t}(0)| = \max_{r \in [0,1]} |(T+1)^{-1/2} y_{i,[r(T+1)]}(0)| \Rightarrow \max_{r \in [0,1]} \nu_i(r)$$

by the continuous mapping theorem.

To show Part (d), we combine (C.2) and (C.4), and have

$$\|D_T(\widehat{\beta}_i - \beta_i)\| = \left\| \begin{bmatrix} \widehat{a}_i - a_i \\ \sqrt{T}(\widehat{b}_i - b_i) \end{bmatrix} \right\| = o_p(1).$$

Therefore,

$$\|(\widehat{\theta} - \theta_0) D_T\|_F = \sqrt{\sum_{i=1}^{N} \|D_T(\widehat{\beta}_i - \beta_i)\|^2} = o_p(1).$$

The second half of Part (d) is also satisfied since $\widehat{\theta}^{(t)} = \widehat{\theta}$ for each t.

Part (e) is assumed and Part (f) is trivial for $W_T = I$. □

Proof of Theorem 3.3. We use similar strategy as we do in the proof of Theorem 3.2. Let

$$L_{1,T}(\epsilon) = \left\{ \|(\widehat{\theta} - \theta_0) D_T\|_F \le \epsilon, \max_{t=1,\dots,T} \|(\widehat{\theta}^{(t)} - \theta_0) D_T\|_F \le \epsilon \right\},$$
$$L_{2,T}(c) = \left\{ \max_{t \le T+1} \|D_T^{-1} x_t\| \le c \right\},$$
$$L_{3,T}(\eta) = \left\{ \|\widehat{G}' C' W_T C \widehat{G} - G' C' W C G\|_F < \eta \right\}.$$

By Assumption 3.3(d), there exists a positive sequence $\{\epsilon_T\}_{T \ge 1}$ such that $\epsilon_T \to 0$ and $\Pr(L_{1,T}(\epsilon_T)) \to 1$. Let $c_T = 1/\sqrt{\epsilon_T}$. So we have $c_T \to \infty$ and $c_T \epsilon_T \to 0$. By

Assumption 2(c), we must have $\Pr(L_{2,T}(c_T)) \to 1$. By Assumption 3.1(c) and Assumption 3.2(f), there exists a positive sequence $\{\eta_T\}_{T\geq 1}$ such that $\eta_T \to 0$ and $\Pr(L_{3,T}(\eta_T)) \to 1$. Let $L_T = L_{1,T}(\epsilon_T) \cap L_{2,T}(c_T) \cap L_{3,T}(\eta_T)$, then we have $\Pr(L_T) \to 1$ and $\Pr(L_T^c) \to 0$.

Suppose L_T holds. Then, for some $\theta = \widehat{\theta}$ or $\theta = \widehat{\theta}^{(t)}$ and for some $t = 1, \ldots, T$, we have

$$|\widehat{P}_t(\theta) - P_t(\theta_0)| \leq |\widehat{P}_t(\theta) - P_t(\theta)| + |P_t(\theta) - P_t(\theta_0)|. \tag{C.10}$$

Note that

$$\begin{aligned}
|\widehat{P}_t(\theta) - P_t(\theta)| &= \left|(Y_t - \theta x_t)'(\widehat{G}'C'W_TC\widehat{G}) - G'C'WCG)(Y_t - \theta x_t)\right| \\
&\leq \|Y_t - \theta x_t\|^2 \|(\widehat{G}'C'W_TC\widehat{G} - G'C'WCG\|_F \\
&\leq \|u_t + (\theta_0 - \theta)x_t\|^2 \cdot \eta_T \\
&\leq (\|u_t\| + \|(\theta_0 - \theta)D_TD_T^{-1}x_t\|)^2\eta_T \\
&\leq (\|u_t\| + \|(\theta_0 - \theta)D_T\|_F\|D_T^{-1}x_t\|)^2\eta_T \\
&\leq (\|u_t\| + \epsilon_T c_T)^2\eta_T
\end{aligned} \tag{C.11}$$

and

$$\begin{aligned}
|P_t(\theta) - P_t(\theta_0)| &= |(Y_t - \theta x_t)'G'C'WCG(Y_t - \theta x_t) - (Y_t - \theta_0 x_t)'G'C'WCG(Y_t - \theta_0 x_t)| \\
&\leq |(Y_t - \theta x_t)'G'C'WCG(Y_t - \theta x_t) - (Y_t - \theta x_t)'G'C'WCG(Y_t - \theta_0 x_t)| \\
&\quad + |(Y_t - \theta x_t)'G'C'WCG(Y_t - \theta_0 x_t) - (Y_t - \theta_0 x_t)'G'C'WCG(Y_t - \theta_0 x_t)| \\
&= |(u_t + (\theta_0 - \theta)x_t)'G'C'WCG(\theta_0 - \theta)x_t| + |((\theta_0 - \theta)x_t)'G'C'WCGu_t| \\
&\leq \|u_t + (\theta_0 - \theta)D_TD_T^{-1}x_t\|\|G'C'WCG\|_F\|(\theta_0 - \theta)D_TD_T^{-1}x_t\| \\
&\quad + \|(\theta_0 - \theta)D_TD_T^{-1}x_t\|\|G'C'WCG\|_F\|u_t\| \\
&\leq (\|u_t\| + \epsilon_T c_T)\|G'C'WCG\|_F\epsilon_T c_t + \epsilon_T c_T\|G'C'WCG\|_F\|u_t\| \\
&= (2\|u_t\| + \epsilon_T c_T)\|G'C'WCG\|_F\epsilon_T c_t.
\end{aligned} \tag{C.12}$$

Combining (C.10), (C.11), and (C.12), we have

$$|\widehat{P}_t(\theta) - P_t(\theta_0)| \leq g(\epsilon_T, c_T, \eta_T),$$

where

$$g_t(\epsilon_T, c_T, \eta_T) = (\|u_t\| + \epsilon_T c_T)^2 \eta_T + (2\|u_t\| + \epsilon_T c_T)\|G'C'WCG\|_F \epsilon_T c_t.$$

By Assumption 3.1(a), $g_t(\epsilon_T, c_T, \eta_T)$ is identically distributed across t for a fixed T.

To show part (a), note that under null,

$$\begin{aligned}
P &= (C\widehat{\alpha} - d)' W_T (C\widehat{\alpha} - d) \\
&= (C(\alpha + Gu_{T+1} + o_p(1)) - d)'(W + o_p(1))(C(\alpha + Gu_{T+1} + o_p(1)) - d) \\
&= (CGu_{T+1} + o_p(1))'(W + o_p(1))(CGu_{T+1} + o_p(1)) \\
&= u'_{T+1} G'C'WCGu_{T+1} + o_p(1).
\end{aligned}$$

The second equality is by Theorem 3.1. Since $P_\infty = u'_1 G'C'WCGu_1$, we have $P \to_d P_\infty$ by stationary of $\{u_t\}_{t \geq 1}$.

Part (b)-(d) can be shown using the same strategy as in the proof of Theorem 3.2, with $g_t(\epsilon_T, c_T, \eta_T)$ in place of $g_t(\epsilon_T, c_T)$, and θ in place of β, so is omitted here. □

REFERENCES

Abadie, A. and M. D. Cattaneo (2018). Econometric Methods for Program Evaluation. *Annual Review of Economics 10*(1), 465–503.

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association 105*(490), 493–505.

Abadie, A. and J. Gardeazabal (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review 93*(1), 113–132.

Abramowitz, M. and I. A. Stegun (1965). *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. Dover Publications.

Anderson, T. W. and H. Rubin (1949). Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *The Annals of Mathematical Statistics 20*(1), 46–63.

Andrews, D. W. K. (2003). End-of-Sample Instability Tests. *Econometrica 71*(6), 1661–1694.

Andrews, D. W. K., M. J. Moreira, and J. H. Stock (2006). Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression. *Econometrica 74*(3), 715–752.

Andrews, I. and T. B. Armstrong (2017). Unbiased instrumental variables estimation under known first-stage sign. *Quantitative Economics 8*(2), 479–503.

Andrews, I. and A. Mikusheva (2016). Conditional Inference With a Functional Nuisance Parameter. *Econometrica 84*(4), 1571–1612.

Andrews, I., J. H. Stock, and L. Sun (2019). Weak Instruments in Instrumental Variables Regression: Theory and Practice. *Annual Review of Economics 11*(1), 727–753.

Assoud, P. (1977). *Espaces Métriques, Plongements, Facteurs*. Doctoral Dissertation, Université de Paris XI, 91405 Orsay France.

Basse, G., A. Feller, and P. Toulis (2017). Conditional randomization tests of causal effects with interference between units. *arXiv preprint arXiv:1709.08036*.

Bentkus, V., M. Bloznelis, and F. Götze (1996). A Berry-Esséen bound for student's statistic in the non-i.i.d. case. *Journal of Theoretical Probability 9*(3), 765–796.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How Much Should We Trust Differences-In-Differences Estimates? *The Quarterly Journal of Economics 119*(1), 249–275.

Bester, C. A., T. G. Conley, and C. B. Hansen (2011a). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics 165*(2), 137–151.

Bester, C. A., T. G. Conley, and C. B. Hansen (2011b). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics 165*(2), 137 – 151.

Bester, C. A., T. G. Conley, C. B. Hansen, and T. J. Vogelsang (2008). Fixed-b asymptotics for spatially dependent robust nonparametric covariance matrix estimators. Mimeo.

Bolthausen, E. (1982, 11). On the central limit theorem for stationary mixing random fields. *Ann. Probab. 10*(4), 1047–1050.

Cameron, A. C. and D. L. Miller (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources 50*(2), 317–372.

Canay, I. A., J. P. Romano, and A. M. Shaikh (2017). Randomization Tests Under an Approximate Symmetry Assumption. *Econometrica 85*(3), 1013–1030.

Cao, J., C. Hansen, D. Kozbur, and L. Villacorta (2019). Inference for Dependent Data with Cluster Learning. *Working paper*.

Cavallo, E., S. Galiani, I. Noy, and J. Pantano (2013). Catastrophic Natural Disasters and Economic Growth. *Review of Economics and Statistics 95*(5), 1549–1561.

Chernozhukov, V., K. Wuthrich, and Y. Zhu (2017). An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls. *arXiv preprint arXiv:1712.09089*.

Coibion, O., Y. Gorodnichenko, and D. Koustas (2017). Consumption Inequality and the Frequency of Purchases. *American Economic Journal: Macroeconomics, forthcoming*.

Condra, L. N., J. D. Long, A. C. Shaver, and A. L. Wright (2018). The Logic of Insurgent Electoral Violence. *American Economic Review 108*(11), 3199–3231.

Conley, T. G. (1996). *Econometric Modelling of Cross-Sectional Dependence*. Ph.D. Dissertation, University of Chicago.

Conley, T. G. (1999). Gmm estimation with cross sectional dependence. *Journal of Econometrics 92*, 1–45.

Conley, T. G. and B. Dupor (2003). A spatial analysis of sectoral complementarity. *Journal of Political Economy 111*(2), 311–352.

Conley, T. G. and C. R. Taber (2011). Inference with "Difference in Differences" with a Small Number of Policy Changes. *Review of Economics and Statistics 93*(1), 113–125.

Conley, T. G. and G. Topa (2002). Socio-economic distance and spatial patterns in unemployment. *Journal of Applied Econometrics 17*(4), 303–327.

Dell, M. (2012). Path Dependence in Development: Evidence from the Mexican Revolution. *Working paper*.

Deryugina, T., G. Heutel, N. H. Miller, D. Molitor, and J. Reif (2019). The mortality and medical costs of air pollution: Evidence from changes in wind direction. *American Economic Review 109*(12), 4178–4219.

Doudchenko, N. and G. W. Imbens (2016). Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis. *arXiv preprint arXiv:1610.07748*.

Fama, E. F. and J. MacBeth (1973a). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy 81*, 607–636.

Fama, E. F. and J. D. MacBeth (1973b). Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy 81*(3), 607–636.

Fang, Z. and A. Santos (2018). Inference on Directionally Differentiable Functions. *The Review of Economic Studies 86*(1), 377–412.

Ferman, B. (2019). A simple way to assess inference methods.

Ferman, B. and C. Pinto (2017). Placebo Tests for Synthetic Controls. Working paper.

Ferman, B. and C. Pinto (2019a). Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity. *The Review of Economics and Statistics 101*(3), 452–467.

Ferman, B. and C. Pinto (2019b). Synthetic Controls with Imperfect Pre-Treatment Fit. Working paper.

Firpo, S. and V. Possebom (2018). Synthetic Control Method: Inference, Sensitivity Analysis and Confidence Sets. *Journal of Causal Inference 6*(2).

Hagemann, A. (2019a). Permutation inference with a finite number of heterogeneous clusters. *arXiv preprint arXiv:1907.01049*.

Hagemann, A. (2019b). Placebo inference on treatment effects when the number of clusters is small. *Journal of Econometrics 213*(1), 190–209.

Hahn, J. and R. Shi (2017). Synthetic Control and Inference. *Econometrics 5*(4), 52.

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

Hansen, B. E. and S. Lee (2019). Asymptotic theory for clustered samples. *Journal of Econometrics 210*(2), 268–290.

Hansen, C. B. (2007). Asymptotic properties of a robust variance matrix estimator for panel data when T is large. *Journal of Econometrics 141*(2), 597–620.

Harari, M. (2020). Cities in Bad Shape: Urban Geometry in India. *American Economic Review 110*(8), 2377–2421.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

Hayashi, F. (2000). *Econometrics*. Princeton University Press.

Ibragimov, R. and U. K. Müller (2010). t-Statistic Based Correlation and Heterogeneity Robust Inference. *Journal of Business & Economic Statistics 28*(4), 453–468.

Jenish, N. and I. Prucha (2007). Central limit theorems and uniform laws of large numbers for arrays of random fields. Mimeo.

Jenish, N. and I. R. Prucha (2009, may). Central Limit Theorems and Uniform Laws of Large Numbers for Arrays of Random Fields. *Journal of econometrics 150*(1), 86–98.

Jirak, M. (2016, 05). Berry–esseen theorems under weak dependence. *Ann. Probab. 44*(3), 2024–2063.

Kaji, T. (2021). Theory of Weak Identification in Semiparametric Models. *Econometrica 89*(2), 733–763.

Kelejian, H. H. and I. Prucha (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review 40*, 509–533.

Kelejian, H. H. and I. Prucha (2001). On the asymptotic distribution of the moran i test statistic with applications. *Journal of Econometrics 104*, 219–257.

Kleibergen, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica 70*(5), 1781–1803.

Kreif, N., R. Grieve, D. Hangartner, A. J. Turner, S. Nikolova, and M. Sutton (2016). Examination of the Synthetic Control Method for Evaluating Health Policies with Multiple Treated Units. *Health Economics 25*(12), 1514–1528.

Lazarus, E., D. J. Lewis, and J. H. Stock (2019). The size-power tradeoff in har inference. *SSRN Working Paper 165*(2), 137 – 151.

Lazarus, E., D. J. Lewis, J. H. Stock, and M. W. Watson (2018). HAR Inference: Recommendations for Practice. *Journal of Business & Economic Statistics 36*(4), 541–559.

Lee, D. L., J. McCrary, M. J. Moreira, and J. Porter (2020). Valid t-ratio Inference for IV.

Lee, L.-f. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial econometric models. *Econometrica 72*, 1899–1926.

Lee, L.-f. (2007a). Gmm and 2sls estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics 137*, 489–514.

Lee, L.-f. (2007b). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics 140*, 333–374.

Li, K. T. (2019). Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods. *Journal of the American Statistical Association*, 1–40.

MacKinnon, J. G. and M. D. Webb (2017). Wild Bootstrap Inference for Wildly Different Cluster Sizes. *Journal of Applied Econometrics 32*(2), 233–254.

Mills, B. (2019). Inference Under First-Stage Sign Information in the Instrumental Variables Model. *Working paper*.

Moreira, H. and M. J. Moreira (2019). Optimal two-sided tests for instrumental variables regression with heteroskedastic and autocorrelated errors. *Journal of Econometrics 213*(2), 398–433.

Moreira, M. J. (2003). A Conditional Likelihood Ratio Test for Structural Models. *Econometrica 71*(4), 1027–1048.

Mueller, U. and M. Watson (2014). Spatial correlation robust inference. *Working paper*.

Newey, W. K. and D. McFadden (1994). Chapter 36 Large sample estimation and hypothesis testing. *Handbook of Econometrics 4*, 2111–2245.

Newey, W. K. and K. D. West (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica 55*(3), 703.

Pesaran, M. H., M. H. Pesaran, Y. Shin, and R. P. Smith (1999). Pooled Mean Group Estimation of Dynamic Heterogeneous Panels. *Journal of the American Statistical Association 94*(446), 621–634.

Pesaran, M. H. and R. Smith (1995). Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics 68*(1), 79–113.

Robbins, M. W., J. Saunders, and B. Kilmer (2017). A Framework for Synthetic Control Methods With High-Dimensional, Micro-Level Data: Evaluating a Neighborhood-Specific Crime Intervention. *Journal of the American Statistical Association 112*(517), 109–126.

Rosenbaum, P. R. (2007). Interference Between Units in Randomized Experiments. *Journal of the American Statistical Association 102*(477), 191–200.

Rudelson, M. and R. Vershynin (2008). The Littlewood-Offord problem and invertibility of random matrices. *Advances in Mathematics 218*(2), 600 – 633.

Staiger, D. and J. H. Stock (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica 65*(3), 557–586.

Sun, Y. and M. S. Kim (2015). Asymptotic F-Test in a GMM Framework with Cross-Sectional Dependence. *Review of Economics and Statistics 97*(1), 210–223.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Vazquez-Bare, G. (2017). Identification and Estimation of Spillover Effects in Randomized Experiments. *arXiv preprint arXiv:1711.02745*.

Voinov, V. G. and M. S. Nikulin (1993). *Unbiased Estimators and Their Applications, Vol. 1: Univariate Case*. Dordrecht: Kluwer Academic Publishers.

Winkelbauer, A. (2012). Moments and Absolute Moments of the Normal Distribution. *arXiv preprint arXiv:1209.4340*.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (Second ed.). Cambridge, Massachusetts: The MIT Press.

Xu, Y. (2017). Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models. *Political Analysis 25*(01), 57–76.

Young, A. (2019). Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *Quarterly Journal of Economics 134*(2), 557–598.

Yu, M., V. Gupta, and M. Kolar (2019). Constrained High Dimensional Statistical Inference. *arXiv preprint arXiv:1911.07319*.

Zarantonello, E. H. (1971). Projections on Convex Sets in Hilbert Space and Spectral Theory. In *Contributions to Nonlinear Functional Analysis*, pp. 237–424. Elsevier.