Inference with Correlated Clusters

David Powell

RAND Labor & Population

WR-1137-1

May 2017

This paper series made possible by the NIA funded RAND Center for the Study of Aging (P30AG012815) and the RAND Labor and Population Unit.

RAND working papers are intended to share researchers' latest findings and to solicit informal peer review. They have been approved for circulation by RAND Labor and Population but have not been formally edited or peer reviewed. Unless otherwise indicated, working papers can be quoted and cited without permission of the author, provided the source is clearly referred to as a working paper. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors. RAND® is a registered trademark.

For more information on this publication, visit www.rand.org/pubs/working_papers/WR1137-1.html

Published by the RAND Corporation, Santa Monica, Calif.

© Copyright 2017 RAND Corporation

RAND® is a registered trademark

Limited Print and Electronic Distribution Rights

This document and trademark(s) contained herein are protected by law. This representation of RAND intellectual property is provided for noncommercial use only. Unauthorized posting of this publication online is prohibited. Permission is given to duplicate this document for personal use only, as long as it is unaltered and complete. Permission is required from RAND to reproduce, or reuse in another form, any of its research documents for commercial use. For information on reprint and linking permissions, please visit www.rand.org/pubs/permissions.html.

The RAND Corporation is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Support RAND

Make a tax-deductible charitable contribution at
www.rand.org/giving/contribute

www.rand.org

Panel Data Inference with Dependent Clusters^{*}

David Powell[†]

First Version: June 2015

This Version: May 2017

Abstract

This paper introduces a method which permits valid inference given a finite number of heterogeneous, correlated clusters. Many inference methods assume clusters are asymptotically independent or model dependence across clusters as a function of a distance metric. With panel data, these restrictions are unnecessary. This paper relies on a test statistic using the mean of the cluster-specific scores normalized by the variance and simulating the distribution of this statistic. To account for cross-cluster dependence, the relationship between each cluster is estimated, permitting the independent component of each cluster to be isolated. The method is simple to implement, can be employed for linear and nonlinear estimators, places no restrictions on the strength of the correlations across clusters, and does not require prior knowledge of which clusters are correlated or even the existence of independent clusters. In simulations, the procedure rejects at the appropriate rate even in the presence of highly-correlated clusters.

Keywords: Finite Inference, Correlated Clusters, Fixed Effects, Panel Data, Hypothesis Testing, Score Statistic

JEL classification: C12, C21, C23, C33

^{*}This paper benefited from comments received at several conferences and seminars, especially those from discussants Pravin Trivedi and Austin Nichols. I also had helpful discussions with Abby Alpert, Anirban Basu, and Colin Cameron.

[†]RAND, dpowell@rand.org

1 Introduction

It has become common in applied work to account for within-cluster dependence while assuming that the cluster themselves are independent. Most inference methods require that clusters are independent, at least asymptotically, or that any such dependence is a function of a distance metric. This paper introduces an inference procedure that is valid for a fixed number of heterogeneous, correlated clusters. The dependence across clusters is not pre-specified as a function of a distance measure and there are no restrictions on the strength of the dependence. Many empirical applications rely on panel data and there is widespread interest in appropriate inference in such cases (Bertrand, Duflo and Mullainathan (2004)). The assumption of independence across clusters can be relaxed, and the proposed method permits within-cluster dependence as well as cross-cluster dependence. There are many reasons that clusters may be correlated and failing to account for dependence across clusters can severely distort inference. With panel data, it is possible to estimate the dependence across clusters and isolate the independent components of each cluster. Given these independent components, appropriate inference is possible. The proposed inference method is straightforward to implement, can be used for linear and nonlinear estimation, and is valid for small N . It does not require any prior knowledge concerning which clusters are possibly correlated and even permits proper inference when there are no independent clusters.

It is typical to account for dependence within “clusters” which are often the same as the units (e.g., “clustering by state”) such that $G = N$, where N is the number of units and G is the number of clusters. Given the popularity of adjusting for clustering at the unit level, I will use “unit” and “cluster” interchangeably, unless otherwise specified when discussing the approach of including multiple units per cluster ($G < N$). Units may be correlated for several reasons such as spatial proximity, similar political institutions, and similarities on unobserved dimensions. Moreover, one motivation for the approach in this paper is the popularity of panel data models which include unit and time fixed effects. The inclusion of time fixed effects in linear models is equivalent to de-meaning the outcome and explanatory variables within each year. This de-meaning may induce dependence across units. With panel data, it is unnecessary to pre-determine which units are correlated or parameterize the source of dependence (e.g., spatial proximity). Instead, this dependence can be estimated empirically by studying co-movements in the scores to isolate the independent variation within each cluster for appropriate inference.

The recognition of dependence in applied work has motivated estimation of covariance matrices which model the nature of this dependence. In the time series context, Newey and West (1987) introduced a “plug-in” covariance matrix for dependence over time. Conley (1999) discussed an equivalent procedure for cross-sectional data with dependence modeled as a function of “economic distance.” It has become common to use a cluster covariance matrix estimator (CCE) which permits arbitrary within-cluster dependence (Liang and Zeger (1986); Arellano (1987)). As discussed in MacKinnon and Webb (2016), general results on CCE estimation assume that the number of clusters approaches infinity, the within-cluster error correlations are constant across clusters, each cluster is the same size, and the clusters are independent. The inference procedure introduced in this paper will not require these assumptions. This paper focuses on inference when N is small and T is large given recent interest in this case (e.g., Bertrand, Duflo and Mullainathan (2004); Hansen (2007); Ibragimov and Müller (2010); Bester, Conley and Hansen (2011)), but the method applies equally well to the large N , small T case as discussed in Section 4.3.1 and there is little loss in assuming large T , instead of large N , for the main discussion of the paper.

While many inference procedures require independence across clusters, recent work introduces restrictions and methods in which weak dependence – which converges to zero – is permitted. Bester, Conley and Hansen (2011) (BCH) analyzes conditions under which CCE can be used to construct appropriate t and Wald statistics while permitting weak dependence across clusters. This method builds on the “fixed- b ” asymptotics discussed in Kiefer and Vogelsang (2002, 2005). Modeling dependence across observations as a function of proximity in space and time (or other dimensions), BCH notes that as the size of the clusters grows, most observations are far from each other on these dimensions. The approaches proposed in Ibragimov and Müller (2010) and Canay et al. (forthcoming), discussed below, also permit finite dependence across clusters. Instead of weak dependence, the inference procedure discussed in this paper places no restrictions on the strength of the dependence across clusters and does not require it to converge to zero as the number of clusters grows or the clusters grow in size.

One way to account for correlations across units is to include correlated units within the same cluster such that $N > G$. This approach requires (1) knowledge of which units are independent and (2) the existence of asymptotically independent clusters. The proposed inference procedure exploits the insight that it is not necessary to have multiple independent units. Each unit has some independent component which can be extracted and subsequently used for proper inference, permitting dependence across all observations.

This paper develops a correlation-adjusted inference (CAI) procedure which uses the scores of extremum estimators. The method employs a Wald test statistic and simulates the distribution of this test statistic under the null hypothesis by perturbing the scores. With asymptotically independent¹ clusters, the simulated distribution can be used to generate valid p-values. CAI uses the panel nature of the data to estimate the relationship between the clusters, permitting correlations across all clusters. Given these estimates, it is possible to create N asymptotically (as $T \rightarrow \infty$) independent functions using only the independent components of the scores for each cluster. Given asymptotically uncorrelated scores, use of the Wald test statistic and its simulated counterparts is appropriate.

This paper makes two primary contributions. First, it introduces a useful inference procedure in the presence of asymptotically independent clusters which is straightforward to implement for linear and nonlinear models. The procedure is computationally attractive and valid for a small number of clusters (or small T) while permitting within-cluster dependence. It imposes no restrictions on homogeneity across clusters and does not require the full model to be estimable for each cluster by itself. These properties distinguish the inference procedure from many approaches recently introduced in the literature.

Second, the inference procedure is simple to extend to permit valid inference in the presence of correlated clusters. There are no restrictions on the strength of this dependence. The method uses the panel nature of the data to estimate the relationship between the scores of the clusters and then generates independent functions for each cluster. The procedure does not require prior knowledge about which clusters are possibly correlated but, instead, empirically estimates the correlations across all clusters. In one set of simulations, I generate highly-correlated clusters which lead to rejection rates by existing methods of over 50%. CAI rejects at appropriate rates. This procedure is simple to implement as estimating the dependence across clusters only requires OLS estimation.

In the next section, I further discuss the finite inference literature and motivate the need to relax assumptions of independence or weak dependence across clusters. Section 3 discusses the general framework and introduces the inference procedure for asymptotically independent clusters. While a primary motivation of this paper is to introduce an inference procedure which is valid for correlated clusters, this section will discuss the merits of the proposed approach even when the clusters are independent. Section 4 considers correlated

¹Because the inference procedure relies on functions which converge to normally-distributed random variables, I use “independent” and “uncorrelated” interchangeably throughout the paper.

clusters and extends the method by creating independent functions from correlated clusters. It also discusses extensions such as the large N , small T case. Section 5 includes several sets of simulations and compares CAI to other inference methods in the literature. I also apply the CAI method to study the employment consequences of state-level minimum wage increases. I conclude in Section 6.

2 Background

2.1 Literature

Many economic applications use dependent data, and proper inference necessitates accounting for this dependence. Presuming independence inappropriately will often make it more likely that one rejects a true null hypothesis. The use of CCE is common in empirical work and it is typical to adjust for clustering based on geographic units such as states and assume that each state represents an independent experiment.

Recent work has considered the limitations of CCE while others have formalized alternative assumptions under which CCE is valid. Hansen (2007) provides conditions under which CCE provides valid inference for fixed N as $T \rightarrow \infty$. Because of the requirement that clusters be independent, the model analyzed in Hansen (2007) includes unit fixed effects but *not* time fixed effects. Carter et al. (forthcoming) considers the effects of cluster-size heterogeneity as well as other sources of heterogeneity on inference, finding that clusters of different sizes can lead to improper inference in finite samples using CCE.

BCH considers assumptions which permit valid inference for a finite number of large clusters that are weakly dependent in finite samples but asymptotically independent. Generating test statistics using CCE under these conditions requires assumptions that the clusters are the same size² and, to generate traditional t - and F -statistics, it is necessary to enforce homogeneity across clusters of both $\mathbf{x}'_g \mathbf{x}_g$ and $\mathbf{x}'_g \boldsymbol{\epsilon}_g$ in the linear case (similar assumptions are required in Hansen (2007)). These assumptions rule out typical difference-in-differences empirical strategies since they would require that each group is treated the same proportion of time as well as many other cases where heterogeneity in these matrices is likely common. Furthermore, fixed effects potentially pose problems in this set-up (see footnote 21 of their paper). For panel data analysis, BCH recommends creating large spatially-grouped

²This assumption is relaxed in Appendix B of their paper to permit different group sizes as long as no group is dominant.

clusters under the assumption that dependence is a function of geographic proximity and time. As an example, they consider a simulated policy which is spatially-correlated across states. They suggest creating larger clusters such as dividing the United States into two clusters³ and using CCE along with a t -distribution with one degree of freedom.

Other work has introduced alternative inference methods for difference-in-differences estimation when the number of clusters is small, including Donald and Lang (2007) and Wooldridge (2003). Conley and Taber (2011) proposed an inference procedure for applications studying the effects of a policy adopted by a small number of states under the assumption that the errors across groups are identically and independently distributed.

Ibragimov and Müller (2010) introduced a method which requires estimating the model separately for each cluster. The method is valid for a small number of heterogeneous, large clusters. The authors suggest partitioning the data into $q \geq 2$ (asymptotically) independent groups and estimating the model by cluster. Under an assumption that a central limit theorem holds for each cluster-specific estimate, this procedure provides q normally distributed variables. Building on a result from Bakirov and Székely (2006), they show that a t -statistic derived from these estimates provides a valid t -test. The test is conservative such that the rejection probability may be well below the statistical significance level. Furthermore, the test is invalid for $q > 14$ or a statistical significance threshold greater than 8.3% and is only appropriate for tests involving a scalar, ruling out joint hypothesis tests. Finally, the procedure requires estimating the model separately for each cluster. In cases where variation across clusters is necessary to estimate the parameters, this requirement may be problematic and often demands that multiple clusters be combined so that the parameters can be identified. Ibragimov and Müller (2016) extends the test to two-sample problems and discusses applications to difference-in-differences estimation.

Canay et al. (forthcoming) builds on the Ibragimov and Müller (2010) approach, using randomization tests to simulate the distribution of the test statistic. Their method permits multiple hypothesis testing and is valid for all significance thresholds and number of clusters. The requirement to estimate all parameters for each cluster remains. Canay et al. (forthcoming) discusses their method given estimates which are symmetrically-distributed around the true parameter under the null hypothesis. In this paper, I provide standard conditions under which the Central Limit Theorem provides asymptotically-normal functions. However, the approach will be valid as long as the functions are symmetric in finite samples

³They test the use of CCE with other sizes as well.

or asymptotically. As in Canay et al. (forthcoming), symmetry can be assumed and the inference approach in this paper is appropriate given this assumption.

While existing work permits weak dependence across clusters (Ibragimov and Müller (2010); Bester et al. (2011); Canay et al. (forthcoming)), this paper permits dependence across clusters without the requirement that this dependence converges to zero as cluster sizes grow. Other work (e.g., Vogelsang (2012); Driscoll and Kraay (1998); Kim and Sun (2013)) has considered cases in which there is both temporal and spatial dependence while modeling at least one dimension of dependence as a function of spatial/temporal proximity using extensions of heteroskedasticity autocorrelation consistent (HAC) robust standard errors. CAI does not require prior knowledge about which clusters are dependent or reliance on a distance metric to parameterize the dependence.

Computationally, the CAI method shares some similarities with the approach introduced by Kline and Santos (2012), a score-based bootstrap which simulates the distribution of the parameters by randomly weighting the score functions. This procedure is offered as an alternative to the wild bootstrap since it has computational advantages given that the entire model does not need to be re-estimated to produce simulated values of the parameters. The perturbed scores are multiplied by the inverse of the Hessian to generate new estimates. Kline and Santos (2012) discuss how this approach provides some benefits in terms of asymptotic refinement properties, though the procedure is not introduced with considerations of finite inference or dependence.

2.2 Motivation

An influential literature has introduced methods to account for within-unit dependence. However, it is still generally assumed that clusters are (asymptotically) independent. This paper is motivated by concerns that this assumption is too restrictive in many contexts. There are two sets of reasons that the assumption of independent clusters is likely violated in some empirical applications.

First, clusters may have similar economic institutions, political institutions, industry compositions, and other traits which lead to correlated policy adoption and economic trends. States with similar demographics may be comparably affected by national policies. Importantly, the sources of the dependence may be numerous and unobservable. The potential for correlations across units has been discussed in the context of geographic proximity

(e.g., Barrios et al. (2012)) and several papers have introduced methods to adjust for spatial dependence (e.g., Bester et al. (2016)). These approaches (as well as other discussed above) model dependence as a function of a distance metric. With panel data, this restriction is unnecessary. This paper introduces an approach which relaxes the necessity to parameterize dependence as a function of distance, permitting strong dependence across clusters that may be “far” on observable distance measures.

Second, it is common in applied work to use a difference-in-differences framework by studying a panel of states over time. This approach requires the inclusion of state and time fixed effects. Time fixed effects, however, induce correlations across clusters. This point is made in Hansen (2007) which proves the appropriateness of clustering for small N when such time fixed effects are excluded. Consider a linear panel data model represented by the specification

$$y_{it} = \alpha_i + \gamma_t + \mathbf{x}'_{it}\boldsymbol{\beta} + \epsilon_{it}. \quad (1)$$

Including time fixed effects (γ_t) is equivalent to transforming the residual:

$$\epsilon_{it} - \frac{1}{N} \sum_{j=1}^N \epsilon_{jt}.$$

As discussed in Hansen (2007), this demeaning creates a correlation across clusters of order $O(\frac{1}{N})$ when T is large. This dependence disappears as $N \rightarrow \infty$ but many empirical applications involve a fixed number of units. Given small N , demeaning potentially induces strong dependence across all units. With time fixed effects and arbitrary within-cluster error dependence, it is possible that there are no observations with independent errors.

In complementary work, Powell (2016) introduces a synthetic control method which, for each unit, uses a weighted average of the other units as a counterfactual. This approach also induces correlations across units and these correlations are potentially strong even when N is large. CAI is valid even in cases in which the empirical approach induces strong dependence.

3 Model

In this section, I formalize the assumptions required for the test statistic introduced below. These assumptions are typical regularity conditions for extremum estimators (see Newey and McFadden (1994)). I consider extremum estimators given their generality. The conditions

below nest many popular linear and nonlinear estimators. I assume there exists $\widehat{Q}(\boldsymbol{\theta})$ such that

$$\widehat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \widehat{Q}(\boldsymbol{\theta}),$$

with $\widehat{Q}(\boldsymbol{\theta}) \equiv \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T Q_{it}(\boldsymbol{\theta})$ representing the sample objective function (suppressing the dependence of this function on the data). Θ is the set of possible parameter values. The null hypothesis is represented by

$$H_0: a(\boldsymbol{\theta}_0) = 0,$$

where $a(\boldsymbol{\theta}_0)$ is $r \times 1$. The method proposed in this paper uses only the restricted estimator, similar to Rao's (1948) score statistic and the Lagrange multiplier statistic (Aitchison and Silvey (1958)) for maximum likelihood estimators. The alternative hypothesis is represented by $H_1: a(\boldsymbol{\theta}_0) \neq 0$. The restricted estimator is defined by

$$\widetilde{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \widehat{Q}(\boldsymbol{\theta}) \quad \text{subject to} \quad a(\boldsymbol{\theta}) = 0. \quad (2)$$

The inference procedure will use the score of the objective function, evaluated at the restricted estimator ($\nabla_{\boldsymbol{\theta}} \widehat{Q}(\widetilde{\boldsymbol{\theta}})$), and relies on the asymptotic normality of this function (multiplied by $\sqrt{NT}$). This property or asymptotic symmetry can be assumed from the start and the appropriateness of the inference method discussed below in Section 3.2.2 follows. In this section, I discuss some standard regularity conditions that imply this property.

3.1 Examples

Before discussing some conditions for the test statistic introduced in this paper, I give examples of three common estimators covered by this framework and the corresponding functions that can be used for inference. The inference method uses a subset of the elements of $\nabla_{\boldsymbol{\theta}} \widehat{Q}(\widetilde{\boldsymbol{\theta}})$, which I will denote $\widehat{g}(\widetilde{\boldsymbol{\theta}})$, since some elements of $\nabla_{\boldsymbol{\theta}} \widehat{Q}(\widetilde{\boldsymbol{\theta}})$ are equal to zero whether the null hypothesis is true or not; such elements are uninformative. First, ordinary least squares (OLS) chooses the parameters that minimize the sum of the square of the residuals. Second, quantile regression minimizes a check function (Koenker and Bassett (1978)). Third, probit regression is typically implemented through maximum likelihood to estimate binary choice models.

Consider estimation of specification

$$y_{it} = \alpha + \beta_1 x_{it}^{(1)} + \beta_2 x_{it}^{(2)} + \epsilon_{it},$$

under the null hypothesis $H_0 : \beta_1 = 0$. Let $\mathbf{x}'_{it}\tilde{\boldsymbol{\beta}} = \tilde{\alpha} + \tilde{\beta}_2 x_{it}^{(2)}$, where $\tilde{\boldsymbol{\beta}}$ represents the restricted estimates. The score function for unit i under the null hypothesis is

$$\text{OLS: } \hat{g}_i(\tilde{\boldsymbol{\beta}}) = \frac{1}{T} \sum_{t=1}^T x_{it}^{(1)} (y_{it} - \mathbf{x}'_{it}\tilde{\boldsymbol{\beta}})$$

Other elements of the score such as $\frac{1}{T} \sum_{t=1}^T x_{it}^{(2)} (y_{it} - \mathbf{x}'_{it}\tilde{\boldsymbol{\beta}})$ and $\frac{1}{T} \sum_{t=1}^T (y_{it} - \mathbf{x}'_{it}\tilde{\boldsymbol{\beta}})$ are equal to zero, on average, whether or not the null hypothesis is true. For the nonlinear estimators, equivalent functions are:

$$\text{Quantile Regression: } \hat{g}_i(\tilde{\boldsymbol{\beta}}) = \frac{1}{T} \sum_{t=1}^T x_{it}^{(1)} \left[\mathbf{1}(y_{it} \leq \mathbf{x}'_{it}\tilde{\boldsymbol{\beta}}) - \tau \right]$$

$$\text{Probit Regression: } \hat{g}_i(\tilde{\boldsymbol{\beta}}) = \frac{1}{T} \sum_{t=1}^T x_{it}^{(1)} \frac{\phi(\mathbf{x}'_{it}\tilde{\boldsymbol{\beta}})(y_{it} - \Phi(\mathbf{x}'_{it}\tilde{\boldsymbol{\beta}}))}{\Phi(\mathbf{x}'_{it}\tilde{\boldsymbol{\beta}})(1 - \Phi(\mathbf{x}'_{it}\tilde{\boldsymbol{\beta}}))}$$

Most simulations in Section 5 will employ OLS, but I will also include simulations using median regression.

3.2 Inference with Asymptotically Uncorrelated Clusters

Initially, I discuss the inference procedure for asymptotically uncorrelated clusters. Accounting for correlations across clusters will be discussed in Section 4 and involves a straightforward adjustment which generates asymptotically independent functions across clusters. Once this adjustment is made, the method discussed in this section can be used. Here, I formalize the assumptions for the procedure. The conditions are standard regularity conditions for many estimators. The goal is to create score functions which converge to mean-zero symmetric random variables. Other conditions which imply this result are possible or, alternatively, the result can be assumed.

3.2.1 Assumptions

First, I assume that the null hypothesis is true as well as conditions regulating the $a(\cdot)$ function:

A1 (*Null Hypothesis*): (i) $a(\boldsymbol{\theta}_0) = 0$ with $a(\cdot)$ continuously differentiable; (ii) $A \equiv \nabla_{\boldsymbol{\theta}} a(\boldsymbol{\theta}_0)$ has rank r .

Next, I assume that the restricted estimates $\tilde{\boldsymbol{\theta}}$ are consistent. The following conditions imply such consistency:

A2 (*Conditions for Consistency*): There exists $Q_0(\boldsymbol{\theta})$ such that (i) $Q_0(\boldsymbol{\theta})$ is uniquely maximized at $\boldsymbol{\theta}_0$; (ii) $\boldsymbol{\Theta}$ is compact; (iii) $Q_0(\boldsymbol{\theta}_0)$ is continuous; (iv) $\hat{Q}(\boldsymbol{\theta})$ converges uniformly in probability to $Q_0(\boldsymbol{\theta})$.

Condition **A2** assumes identification and regularity conditions which imply consistency of the unrestricted parameter estimates. By Theorem 2.1 in Newey and McFadden (1994), $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$. Together, conditions **A1** and **A2** imply $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ (see Appendix A for a brief discussion). More primitive conditions can be provided in place of **A2**. The next conditions govern dependence within each cluster.

A3 (*Within-Cluster Dependence*): For each i ,

(i) $\{\nabla_{\boldsymbol{\theta}} Q_{it}(\boldsymbol{\theta}_0)\}$ is a strong mixing sequence in t with α of size $-\frac{r}{r-2}$, $r > 2$.

(ii) $E |\nabla_{\boldsymbol{\theta}} Q_{it}(\boldsymbol{\theta}_0)|^{2r} < \Delta < \infty$ for all t, h .

(iii) $\text{Var} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla_{\boldsymbol{\theta}} Q_{it}(\boldsymbol{\theta}_0) \right)$ is uniformly positive definite with constant limit V_i .

Alternative conditions are possible as long as they imply that $\frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla_{\boldsymbol{\theta}} Q_{it}(\boldsymbol{\theta}_0) \xrightarrow{d} N(0, V_i)$. Under **A3**, this condition holds by a CLT which permits dependence and non-identical random variables (see Theorem 5.20 in White (2001)). Assumption **A3** places no restrictions on heterogeneity in V_i across clusters. I also assume that the scores across clusters are asymptotically independent. Defining $\nabla_{\boldsymbol{\theta}} Q_i(\boldsymbol{\theta}_0) \equiv \frac{1}{T} \sum_{t=1}^T \nabla_{\boldsymbol{\theta}} Q_{it}(\boldsymbol{\theta}_0)$:

A4 (*Asymptotic Independence*): $\text{Cov}(\nabla_{\boldsymbol{\theta}} Q_i(\boldsymbol{\theta}_0), \nabla_{\boldsymbol{\theta}} Q_j(\boldsymbol{\theta}_0)) \xrightarrow{p} 0$ for all $i \neq j$ as $T \rightarrow \infty$.

This assumption permits finite dependence across clusters. A motivation of this paper is to consider cases where **A4** is not met by the unadjusted scores because of strong dependence across clusters. The adjustment discussed in Section 4 will create functions that satisfy

A4 even in the presence of correlated clusters. Given **A3** and **A4**, it is straightforward to show that $\frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \nabla_{\theta} Q_{it}(\theta_0) \xrightarrow{d} N(0, V)$, where $V = \frac{1}{N} \sum_{i=1}^N V_i$. The test statistic introduced below relies on the asymptotic normality of $\frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \nabla_{\theta} Q_{it}(\tilde{\theta})$, which requires additional conditions:

A5 (*Conditions for Asymptotic Normality*): (i) $\theta_0 \in \text{interior}(\Theta)$; (ii) $\hat{Q}(\theta)$ twice continuously differentiable in a neighborhood $\mathcal{N}$ of θ_0 ; (iii) There exists $J(\theta)$ continuous at θ_0 and $\sup_{\theta \in \mathcal{N}} \|\nabla_{\theta\theta} \hat{Q}(\theta) - J(\theta)\| \xrightarrow{p} 0$ with $J \equiv J(\theta_0)$ nonsingular.

Conditions **A2-A5** imply $\sqrt{NT}(\hat{\theta} - \theta_0)$ converges to a normally-distributed random variable (see Theorem 3.1 in Newey and McFadden (1994)). Given conditions **A1-A5**, $\sqrt{NT}(\tilde{\theta} - \theta_0)$ also converges to a normally-distributed random variable, as shown in Appendix A when discussing Lemma 3.1 below. The inference procedure will rely on simulating a test statistic by perturbing the score functions. The weights satisfy the following conditions:

A6 (*Weights*): Assume i.i.d weights $\{W_i\}_{i=1}^N$ independent of $\nabla_{\theta} Q_i(\theta_0)$ such that for all i : (i) $E[W_i] = 0$; (ii) $E[W_i^2] = 1$; (iii) $E|W_i|^{2r} < \Delta < \infty$.

Given **A1-A5** and weights that satisfy **A6**, the perturbed scores have the same distribution asymptotically as the score itself. Several distributions of weights meeting these criteria have been suggested in the wild bootstrap literature. In this paper, I focus on the Rademacher distribution which is equal to 1 with probability $\frac{1}{2}$ and -1 with probability $\frac{1}{2}$. The Rademacher distribution has the advantage that $E[W_i^3] = 0$ and $E[W_i^4] = 1$ such that this distribution is able to asymptotically match symmetric distributions and provide asymptotic refinement. I discuss the weights more in Section 3.3.

Lemma 3.1. Assume that **A1-A6** hold with $T \rightarrow \infty$. Then,

$$\begin{aligned} \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \nabla_{\theta} Q_{it}(\tilde{\theta}) &\xrightarrow{d} N(0, \tilde{V}) \\ \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T W_i \nabla_{\theta} Q_{it}(\tilde{\theta}) &\xrightarrow{d} N(0, \tilde{V}), \end{aligned}$$

where $\tilde{V} = (AJ^{-1}A')^{-1}AJ^{-1}VJ^{-1}A'(AJ^{-1}A')^{-1}$.

Because W_i and $\nabla_{\boldsymbol{\theta}} Q_i(\boldsymbol{\theta}_0)$ are independent and $E[W_i] = E[\nabla_{\boldsymbol{\theta}} Q_i(\boldsymbol{\theta}_0)] = 0$,

$$\text{Var}(W_i \nabla_{\boldsymbol{\theta}} Q_i(\boldsymbol{\theta}_0)) = \text{Var}(W_i) \text{Var}(\nabla_{\boldsymbol{\theta}} Q_i(\boldsymbol{\theta}_0)) = \text{Var}(\nabla_{\boldsymbol{\theta}} Q_i(\boldsymbol{\theta}_0)).$$

The weights are constant within each cluster, preserving within-cluster covariances. While Lemma 3.1 includes an explicit formula for $\tilde{V}$, the method introduced below does not require knowledge or estimation of $\tilde{V}$. The importance of Lemma 3.1 is that both vectors converge to the same distribution. Because $E[W_i W_j] = 0$, the convergence of the scores and the perturbed scores to the same distribution can only occur if $\nabla_{\boldsymbol{\theta}} Q_i(\tilde{\boldsymbol{\theta}})$ and $\nabla_{\boldsymbol{\theta}} Q_j(\tilde{\boldsymbol{\theta}})$ are asymptotically uncorrelated since any correlation would not be preserved by the weights. The asymptotic independence of $\nabla_{\boldsymbol{\theta}} Q_i(\tilde{\boldsymbol{\theta}})$ and $\nabla_{\boldsymbol{\theta}} Q_j(\tilde{\boldsymbol{\theta}})$ holds due to consistency of $\tilde{\boldsymbol{\theta}}$ and assumption **A4**. Lemma 3.1 is discussed in more detail in the Appendix.

3.2.2 Test Statistic

The test statistic below does not (necessarily) use all elements of $\nabla_{\boldsymbol{\theta}} \hat{Q}(\tilde{\boldsymbol{\theta}})$ but, instead, relies on a subset given that $\tilde{V}$ is $r \times r$. I refer to this vector as $\hat{g}(\tilde{\boldsymbol{\theta}})$, which is $H \times 1$ where $r \leq H$. This vector has the identification condition that for all h ,

$$E \left[\frac{1}{NT} \sum_i^N \sum_t^T g_{ith}(\boldsymbol{\theta}_0) \right] = 0 \quad \text{if and only if} \quad a(\boldsymbol{\theta}_0) = 0,$$

where h indexes the element and $g_{ith}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} Q_{ith'}(\boldsymbol{\theta})$ for some h' . This condition eliminates elements which are equal to zero regardless of whether the null hypothesis is true or not. The test statistic below is valid if elements which are always equal to zero are used, but their inclusion may reduce power. Define $g_i(\boldsymbol{\theta}) \equiv \frac{1}{T} \sum_{t=1}^T g_{it}(\boldsymbol{\theta})$. For inference, the method uses a Wald test statistic defined by:

$$S = \left(\frac{1}{N} \sum_{i=1}^N g_i(\tilde{\boldsymbol{\theta}}) \right)' \hat{\Sigma}(\tilde{\boldsymbol{\theta}})^{-1} \left(\frac{1}{N} \sum_{i=1}^N g_i(\tilde{\boldsymbol{\theta}}) \right), \quad (3)$$

where $\hat{\Sigma}(\tilde{\boldsymbol{\theta}}) = \frac{1}{N-1} \sum_{i=1}^N \left(g_i(\tilde{\boldsymbol{\theta}}) - \left(\frac{1}{N} \sum_{j=1}^N g_j(\tilde{\boldsymbol{\theta}}) \right) \right) \left(g_i(\tilde{\boldsymbol{\theta}}) - \left(\frac{1}{N} \sum_{j=1}^N g_j(\tilde{\boldsymbol{\theta}}) \right) \right)'$. This approach does not require estimation of V_i , permitting arbitrary within-cluster correlations (as long as a CLT still holds). It uses the empirical variance of the $g_i(\tilde{\boldsymbol{\theta}})$ functions across clusters. This test statistic has similarities to the methods proposed in Ibragimov and Müller (2010)

and Canay et al. (forthcoming), which estimate the parameter(s) of interest in each cluster and then create test statistics using the mean and variance of these estimated parameter(s). Here, the model is not estimated separately for each cluster. Instead, the model is estimated once, under the null hypothesis, and the resulting scores for each cluster are used to construct the test statistic. This property is especially advantageous when it is not possible or it is inefficient to separately estimate the parameters in each cluster.

To simulate the distribution of the test statistic, consider weights $W_i^{(k)}$ which satisfy condition **A6**, where k indexes each “draw.”

$$S^{(k)} = \left(\frac{1}{N} \sum_{i=1}^N W_i^{(k)} g_i(\tilde{\boldsymbol{\theta}}) \right)' \hat{\Sigma}^{(k)}(\tilde{\boldsymbol{\theta}})^{-1} \left(\frac{1}{N} \sum_{i=1}^N W_i^{(k)} g_i(\tilde{\boldsymbol{\theta}}) \right), \quad (4)$$

where $\hat{\Sigma}^{(k)}(\tilde{\boldsymbol{\theta}}) = \frac{1}{N-1} \sum_{i=1}^N \left(W_i^{(k)} g_i(\tilde{\boldsymbol{\theta}}) - \left(\frac{1}{N} \sum_{j=1}^N W_j^{(k)} g_j(\tilde{\boldsymbol{\theta}}) \right) \right) \left(W_i^{(k)} g_i(\tilde{\boldsymbol{\theta}}) - \left(\frac{1}{N} \sum_{j=1}^N W_j^{(k)} g_j(\tilde{\boldsymbol{\theta}}) \right) \right)'$.

The scores are perturbed such that the test statistic distribution is simulated without re-estimating the model. These perturbed test statistics have the same asymptotic distribution as S :

Theorem 3.1. *Assume **A1-A6** hold with $T \rightarrow \infty$. Then, S and $S^{(k)}$ converge to the same distribution.*

This result follows from Lemma 3.1 and the continuous mapping theorem. Consequently, it is possible to simulate the distribution of the test statistic using weights satisfying **A6**. A further discussion of Theorem 3.1 is included in Appendix A. One benefit of this approach is that there is no need to estimate V or $\tilde{V}$, which can be difficult for many estimators (e.g., quantile regression variance estimation is known to be sensitive to choice of bandwidth) and dependence structures.

Define the p-value as $\hat{p}$ using the following formula and K simulations of the test statistic:

$$\hat{p} = \frac{1}{K} \sum_{k=1}^K \mathbf{1}(S \leq S^{(k)}) \quad (5)$$

Under the null hypothesis, $g_i(\tilde{\boldsymbol{\theta}})$ is centered around zero such that large values (in magnitude) of S suggest that the null hypothesis is incorrect. $W_i g_i(\tilde{\boldsymbol{\theta}})$ is centered around zero whether the null hypothesis is true or not. Consequently, when the null hypothesis is not true, a large value of S should be rare in the distribution of $S^{(k)}$.

3.3 Weights

The Rademacher distribution satisfies condition **A6** with the additional advantage in finite samples that it matches symmetric distributions. Canay et al. (forthcoming) introduces an inference method using permutation tests given cluster-specific parameter estimates. They develop their method under the assumption that the estimates are symmetrically-distributed around the null hypothesis. An equivalent assumption can be applied here. While I have relied on the Central Limit Theorem to provide functions that converge to normally-distributed random variables, the method introduced in this section is applicable to cases where the functions have a finite symmetric distribution or where the functions converge to any symmetric distribution.

In simulations, I use the Rademacher distribution to generate weights. This distribution is frequently used for the wild bootstrap with a similar motivation (Cameron et al. (2008)). However, Webb (2014) points out that a 2-point distribution limits the number of unique simulated values of the test statistic that it can produce. With 6 clusters, a 2-point distribution generates only 32 ($= 2^{G-1}$) independent test statistics. The p-values are not point-identified and may include large intervals for small G . Webb (2014) suggests a 6-point distribution for the wild bootstrap with a small number of clusters, noting that it is impossible to find a distribution that matches the first 4 moments of the Rademacher distribution. The recommended weight distribution matches the first three moments with a fourth moment equal to $\frac{7}{6}$, reasonably close to 1. Webb (2014) proposes the following 6-point distribution, each with probability equal to $\frac{1}{6}$:

$$-\sqrt{1.5}, \quad -1, \quad -\sqrt{0.5}, \quad \sqrt{0.5}, \quad 1, \quad \sqrt{1.5}$$

For similar reasons as discussed in Webb (2014), CAI will also be limited in the number of unique simulated test statistics that can be generated for small G using the Rademacher distribution. The benefits of using a 6-point distribution here are similar. In simulations, I use the Webb (2014) distribution when $G < 10$ and the Rademacher distribution otherwise.

3.4 Discussion

This section introduced a valid inference procedure given asymptotically independent clusters. The motivation of this paper is to provide proper inference given dependent clusters,

which requires a simple extension discussed in the next section. However, there are merits to using the above approach when the clusters are asymptotically independent and no adjustment is necessary. The method is simple to implement and does not require estimation of the variance-covariance matrix, which may be difficult to estimate for some estimation techniques and within-cluster dependence structures. It only requires one estimation of the model while enforcing the null hypothesis. It does not require estimation of the parameters for each cluster, allowing estimation of common parameters (such as time fixed effects) across clusters. It also does not require additional assumptions concerning the homogeneity across clusters. Finally, it permits finite dependence across clusters. In simulations, I will include rejection rates using the inference procedure introduced in this section without the correlation adjustment discussed in the next section.

4 Dependent Clusters

Most applied work assumes independence across clusters. However, there are many reasons that this assumption may be violated. Some work has recognized that units that are spatially close may experience common shocks such that the units are not independent. This spatial clustering must be parameterized and the determinants of the dependence must be known. Inference procedures typically require prior knowledge of which units are independent and necessitate independence across some units. Furthermore, many estimation techniques may induce correlations across clusters. With panel data, however, it is possible to observe and estimate dependence across clusters, isolate the independent components of each clusters, and then use the inference procedure introduced in Section 3.2.2. I discuss this method here.

4.1 Cross-Cluster Dependence

I consider each condition, indexed by h , separately. There are H conditions. Each score is divided into a dependent component and an independent component.

A7 (*Dependence Across Clusters*): $g_{ith}(\boldsymbol{\theta}_0)$ can be expressed as

$$g_{ith}(\boldsymbol{\theta}_0) = \sum_{j=i+1}^N \left[b_{jh}^i \times g_{jth}(\boldsymbol{\theta}_0) \right] + \mu_{ith}, \quad (6)$$

with $E[g_{jth}(\boldsymbol{\theta}_0)\mu_{ith}] = 0$ and $E[\mu_{ith}\mu_{jsh}] = 0$ for all s, t, h , and $i \neq j$.

For cluster N , $g_{Nth}(\boldsymbol{\theta}_0) = \mu_{Nth}$. The score for cluster i is modeled as a function of the scores for clusters $i + 1$ to N . This specification is simply modeling the score for each cluster as an independent component, μ_{ith} , and a dependent component, $\sum_{j=i+1}^N \left[b_{jh}^i \times g_{jth}(\boldsymbol{\theta}_0) \right]$. b_{jh}^i represents the relationship between clusters i and j for condition h . Define $g_{ith}^i(\boldsymbol{\theta}_0) \equiv (g_{i+1,th}(\boldsymbol{\theta}_0), \dots, g_{Nth}(\boldsymbol{\theta}_0))'$ and $\mathbf{b}_h^i \equiv (b_{i+1,h}^i, \dots, b_{N,h}^i)'$ so that equation (6) can be rewritten as $g_{ith}(\boldsymbol{\theta}_0) = g_{ith}^i(\boldsymbol{\theta}_0)' \mathbf{b}_h^i + \mu_{ith}$, which will be helpful notation in the subsequent analysis.

It is important to discuss the assumptions that are nested in **A7** as well as what is not restricted by this condition. First, modeling the relationship between scores as linear is not restrictive since we are only interested in the covariances between the cluster-specific scores. Second, there are no restrictions in **A7** concerning the strength of the correlations between any clusters. In fact, this framework permits identical clusters (i.e., $g_{ith}(\boldsymbol{\theta}_0) = g_{jth}(\boldsymbol{\theta}_0)$ for all t, h). Third, equation (6) does not impose any restrictions on within-cluster dependence. Fourth, the triangular framework itself is not an assumption. If there is a common component in multiple clusters, it is important to consider that component “independent” for one of the clusters and “dependent” in the other clusters. This implies a triangular framework and equation (6) models the score of cluster i as a function of the scores of clusters $i + 1, \dots, N$. Fifth, this framework does not require prior knowledge about which clusters are potentially correlated with one another.

Equation (6) assumes that lags and leads for other clusters are not *independently* related to $g_{ith}(\boldsymbol{\theta}_0)$. Instead, dependence will be measured by estimating the co-movements in the scores across clusters. Note that this setup still permits strong correlations between $g_{it}(\boldsymbol{\theta}_0)$ and $g_{js}(\boldsymbol{\theta}_0)$ as long as those correlations operate through $g_{jt}(\boldsymbol{\theta}_0)$ or $g_{is}(\boldsymbol{\theta}_0)$. Given dependence within-cluster i and within-cluster j as well as correlations across clusters in each time period, this framework permits a rich level of dependence across observations in clusters i and j . The idea behind CAI is that by eliminating the correlation across clusters i and j within the same period, this approach eliminates dependence across all observations in clusters i and j . Note further that CAI permits finite dependence across $g_{it}(\boldsymbol{\theta}_0)$ and $g_{js}(\boldsymbol{\theta}_0)$ that does not operate through $g_{jt}(\boldsymbol{\theta}_0)$ or $g_{is}(\boldsymbol{\theta}_0)$. In Section 4.3.2, I discuss extensions of this model to include leads and lags, which further relaxes the assumption that correlations must operate through the period t scores.

To illustrate, consider the linear fixed effects case in equation (1) for small N and assume ϵ_{it} and ϵ_{js} are uncorrelated for all s, t , and $i \neq j$. When de-measured by year (and focus-

ing only on the residuals),⁴ there is dependence between $\epsilon_{it} - \frac{1}{N} \sum_{k=1}^N \epsilon_{kt}$ and $\epsilon_{jt} - \frac{1}{N} \sum_{k=1}^N \epsilon_{kt}$. If there is within-cluster dependence, then there will also be dependence between observations in cluster i and j which are not in the same time period.⁵ However, eliminating the period- t dependence will remove dependence across all observations in i from all observations in j . Note that CAI also permits a wide range of cases where ϵ_{it} and ϵ_{js} are correlated. CAI creates two independent sets of scores for these clusters and valid inference can proceed using the method discussed in Section 3.2.2.

The goal is to create functions which are independent across clusters. It is straightforward to estimate equation (6), plugging in $\tilde{\theta}$ as a consistent estimate for θ_0 , for each cluster i and condition h using OLS. Estimation of equation (6) estimates these co-movements and allows one to subtract off the dependent component of the score to isolate the independent component. With the estimates $\hat{b}_{jh}^i$, it is possible to construct $\hat{\mu}_{ith} = g_{ith}(\tilde{\theta}) - g_{th}^i(\tilde{\theta})' \hat{b}_h^i$ and, subsequently,

$$\hat{s}_{ih}(\tilde{\theta}) \equiv \frac{1}{T} \sum_{i=1}^T \hat{\mu}_{ith}. \quad (7)$$

where $\hat{s}_{ih}(\tilde{\theta})$ is the asymptotically-independent score function for condition h . In practice, equation (6) can be estimated with a constant. The constant is the estimate $\hat{s}_{ih}(\tilde{\theta})$. Define $\hat{b}_h^i \equiv \left[\frac{1}{T} \sum_{t=1}^T g_{th}^i(\tilde{\theta}) g_{th}^i(\tilde{\theta})' \right]^{-1} \left[\frac{1}{T} \sum_{t=1}^T g_{th}^i(\tilde{\theta}) g_{ith}(\tilde{\theta}) \right]$, which represent the estimates from a regression of $g_{ith}(\tilde{\theta})$ on $(g_{i+1,th}(\tilde{\theta}), \dots, g_{Nth}(\tilde{\theta}))$.

Theorem 4.1. *Assume **A1-A3**, **A7**, (i) $E|g_{ith}(\theta)|^{(r+\delta)} < \Delta < \infty$ for all i, t, h and for some $\delta > 0$, for θ in some neighborhood of θ_0 ; (ii) $g_{ith}(\theta)$ continuous at θ_0 with probability one for all i, t, h ; (iii) $E \left[\frac{1}{T} \sum_{t=1}^T g_{th}^i(\theta_0) g_{th}^i(\theta_0)' \right]$ is uniformly positive definite. Then,*

$$\text{Cov} \left[\frac{1}{T} \sum_{t=1}^T \left(g_{ith}(\tilde{\theta}) - g_{th}^i(\tilde{\theta})' \hat{b}_h^i \right), \quad \frac{1}{T} \sum_{t=1}^T \left(g_{jth}(\tilde{\theta}) - g_{th}^j(\tilde{\theta})' \hat{b}_h^j \right) \right] \xrightarrow{p} 0 \quad (8)$$

for all h and for all $i \neq j$.

OLS estimation provides consistent estimates of b_{jh}^i such that $\hat{b}_{jh}^i \xrightarrow{p} b_{jh}^i$. The additional

⁴Correlations in the residuals across clusters do not necessarily imply that the scores are correlated, but I focus on the residuals here for illustrative purposes only.

⁵For example, $\epsilon_{is} - \frac{1}{N} \sum_{k=1}^N \epsilon_{ks}$ may be correlated with $\epsilon_{jt} - \frac{1}{N} \sum_{k=1}^N \epsilon_{kt}$ through its correlation with $\epsilon_{it} - \frac{1}{N} \sum_{k=1}^N \epsilon_{kt}$.

restrictions in Theorem 4.1 include standard assumptions for OLS and rule out perfectly identical clusters (i.e., collinearity).⁶ The correlation adjustment simply involves regressing the score for a cluster on the other scores for clusters $j > i$. This is straightforward to implement. Theorem 4.1 is discussed further in Appendix A.

Theorem 4.1 provides asymptotically independent functions. The conditions needed to require these functions to converge to normally-distributed random variables are similar to the conditions discussed in Section 3.2. First, it is necessary that the restricted estimates are asymptotically normal. Condition **A3**(iii) can be replaced by: $\text{Var} \left(\frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \nabla_{\theta} Q_{it}(\theta_0) \right)$ is uniformly positive definite with constant limit V . Condition **A7** replaces **A4**. As before, $\frac{1}{\sqrt{NT}} \sum_i \sum_t \nabla_{\theta} Q_{it}(\tilde{\theta})$ converges to a normally-distributed random variable, which implies that a subset of those conditions, represented by $\frac{1}{\sqrt{NT}} \sum_i \sum_t g_{it}(\tilde{\theta})$, also converges to a normally-distributed random variable. Consequently, $\frac{1}{\sqrt{NT}} \sum_i \sum_t s_{it}(\tilde{\theta})$ converges to a normally-distributed random variable. Thus, the test statistic in equation (3) can be used given functions $\hat{s}_i(\tilde{\theta})$.

4.2 Summary of CAI

The final inference procedure is summarized in the following steps.

1. Estimate $\tilde{\theta}$ using equation (2).
2. Create $g_{it}(\tilde{\theta})$ for each i, t .
3. Regress $g_{ith}(\tilde{\theta})$ on $g_{i+1,th}(\tilde{\theta}), \dots, g_{Nth}(\tilde{\theta})$ and a constant. The estimated constant is $\hat{s}_{ih}(\tilde{\theta})$. Repeat this step for all (i, h) .
4. Calculate S using $\hat{s}_i(\tilde{\theta})$ and equation (3).
5. Simulate distribution of S using weights meeting condition **A6** and equation (4) to generate $S^{(1)}, \dots, S^{(K)}$.
6. Generate p-value using equation (5).

In all simulations, I use $K = 999$.

⁶Perfectly-identical clusters are acceptable in the equation (6) framework but, in practice, only one of the identical clusters should be used for inference.

4.3 Extensions and Modifications

4.3.1 Large N , Small T

This paper discusses inference for large T . However, the procedure is also appropriate for small T and large N . Valid inference follows from considering each time period as a cluster and then adjusting for correlations across time periods (i.e., within units). In the small T case, the equivalent to equation (6) adjusts for within-unit correlations by regressing the score (for each unit) for period t on the scores for periods $t + 1$ to T .⁷ Conceptually, since the CAI approach allows for within-unit dependence and strong correlations across units, there is little difference whether “clusters” are defined by time or unit as long as a CLT holds in the dimension defined as the cluster. I include simulation results in the next section in which clusters are defined by the time dimension.

4.3.2 Lags and Leads

The only restriction implicit in equation (6) is that correlations across clusters act through the scores in the same time period. Because arbitrary within-cluster correlations are allowed (as long as a CLT holds), this assumption permits dependence across all observations. **A7** rules out the case, as an example, where the score for unit i in time period $t - s$ has an independent relationship with unit j in time period t that does not act through the score of unit i in period t (and this correlation does not converge to zero). This assumption can be relaxed by including lagged values (or leads) in equation (6). With lags, I model the scores using $AR(p)$ notation:

$$g_{ith}(\boldsymbol{\theta}_0) = \sum_{s=0}^p \left(b_{i+1,sh}^i \times g_{i+1,t-s,h}(\boldsymbol{\theta}_0) \right) + \cdots + \sum_{s=0}^p \left(b_{Nsh}^i \times g_{N,t-s,h}(\boldsymbol{\theta}_0) \right) + \mu_{ith}. \quad (9)$$

Leads can also be included. The variable p controls the length of the lags and, while constant in the equation above, can vary by cluster. Thus, equation (6) can be relaxed substantially to allow for more flexibility.

⁷This approach does not require an assumption that the correlations across time periods are the same for all units.

4.3.3 Individual-Level Data

This paper has discussed the inference method as if the observations are defined by unit and time. Many applications, even when clusters are defined at the state-level, use more granular data such as county or individual. Given data for individual j in state i at time t , CAI can be used with $g_{it}(\tilde{\theta}) = \frac{1}{M_{it}} \sum_{m=1}^{M_{it}} g_{mit}(\tilde{\theta})$ where M_{it} represents the number of individuals in state i at time t . The inference procedure proceeds as before. Since M_{it} varies across (i, t) , CAI does not require that clusters are the same size.

4.4 Discussion

The primary contribution of this paper is the introduction of an inference method which adjusts for dependence across clusters. While it is unnecessary to pre-determine the source (e.g., spatial) and magnitude of this dependence, the method does require specifying an estimable equation to measure the correlations across clusters. Equation (6) uses co-movements in scores across clusters to determine empirical correlations. Equation (6) does *not* assume that the dependence between clusters is limited to correlations between observations in the same period but that any correlations across clusters operate through observations in the same period.

This approach, however, rules out arbitrary, unknown correlations across observations in different clusters. For example, equation (6) does not account for a relationship between $g_{it}(\tilde{\theta})$ and $g_{js}(\tilde{\theta})$ that does not operate through $g_{is}(\tilde{\theta})$. These correlations can be modeled, but the appropriate lags and leads must be included in equation (9). To the extent that dependence unaccounted for through equation (9) converges to zero, CAI permits valid inference. It is also still possible to create a larger cluster including both i and j , but this requires prior knowledge about which units belong in the same cluster, similar to existing inference methods. The advantage of CAI, however, is that it permits adjustments across clusters, which is new to the literature.

5 Simulations

Given the generality of the CAI procedure, I test it using several different data generating processes. I start by replicating simulations found in the literature and find that CAI works well relative to existing methods. I will also provide simulations with highly-dependent clusters in which conventional methods perform especially poorly. In each set of simulations,

I include rejection rates using CCE, following BCH to create appropriate t -statistics. I provide three CCE rejection rates - (1) assuming each unit is a cluster ($G = N$); (2) using $G = 4$; (3) using $G = 2$. In all cases, I use a t -distribution with $G - 1$ degrees of freedom. For some simulations, it is clear how to assign units into clusters. In other cases, however, I make this assignment randomly (usually without loss of generality since there is no spatial dependence in those simulations). For simulations involving linear models, I also include rejection rates using the wild cluster bootstrap which simulates the distribution of a Wald statistic and imposes the null hypothesis.⁸

I also provide rejection rates using the “randomization test” method introduced in Canay et al. (forthcoming). In cases where time fixed effects are included in the specification, I create $\frac{N}{2}$ clusters, randomly assigning a pair of units into each cluster. I also include rates from the unadjusted version of the inference method introduced in Section 3.2.2, which I will label as “perturbed scores.” Weights are required for the wild bootstrap, randomization test, perturbed score, and CAI methods. When $G \geq 10$, I use weights generated by the Rademacher distribution. When $G < 10$, I will use the Webb (2014) weights that were motivated by use of the wild bootstrap in small G cases but are also applicable for other approaches which require perturbations.

5.1 BCH Replication

BCH uses monthly state-level unemployment rates from the Bureau of Labor Statistics (BLS) and generates placebo policies which are spatially correlated. They use data from 1976 to 2007 for all states (including Washington DC) except for Alaska and Hawaii and model the log of the unemployment rate as

$$\ln(y_{st}) = \beta x_{st} + \alpha_s + \alpha_t + \epsilon_{st}, \quad (10)$$

where y_{st} is the unemployment rate for state s at time t . The specification includes state and time fixed effects. They set $\beta = 0$. x_{st} is a simulated treatment variable which is generated

⁸Based on findings in Cameron et al. (2008), I implement the “Wild cluster bootstrap- t with H_0 imposed” method. See Appendix B of their paper.

in the following manner:

$$x_{st} = \sigma \left(u_{st} + \gamma \sum_{d(s,r)=1} u_{rt} \right),$$

$$\text{where } u_{st} = \sum_{j=1}^p \rho_j u_{s(t-j)} + \nu_{st},$$

$$\nu_{st} \sim N(0, 1),$$

$d(s, r)$ is equal to 1 for adjacent states s and r (0 otherwise). Spatial and temporal dependence are governed by γ and ρ_j , respectively. I simulate policies in the same manner, using the same data.

Following BCH, I use $\gamma = 0.8$ and set σ such that the standard deviation of x_{st} and $\ln(y_{st})$ are the same. BCH considers two data generating processes. First, they use an AR(13) process with $\rho_1 = 0.95$, $\rho_{12} = 0.68$, and $\rho_{13} = -0.68$. Second, they adopt an AR(1) process with $\rho_1 = 0.8$. I report rejection rates for 5% level tests using CCE for 49 clusters (all units), 4 clusters (states divided into Census regions),⁹ and 2 clusters (dividing the country into two).¹⁰ In each case, I use a t -distribution with $G - 1$ degrees of freedom where G is the number of clusters.

BCH primarily uses first-differences to estimate (10) with only one set of estimates using state and time fixed effects. The inclusion of fixed effects can potentially violate the assumptions permitting use of CCE (see their discussion in footnote 21 for more details). Fixed effects do not pose a problem for CAI. I present results using both first-differences and fixed effects.

The results are reported in Table 1. The first column includes the rejection rates for the AR(13) process, using first-differences for estimation. Clustering by state – typical in applied work – overrejects at a rate of 0.169 due to spatial dependence in x (by design) and the outcome variable (due to correlated shocks in the unemployment rate). The wild bootstrap similarly overrejects. The randomization approach rejects about 10% of the time and the perturbed score method rejects in 14% of simulations. Dividing the states into 4 clusters improves the rejection rate with CCE. Dividing all states into 2 clusters further

⁹I follow BCH in reassigning some states to make the clusters approximately equal-sized. See BCH footnote 23.

¹⁰Again, I follow BCH to assign states to East or West. See BCH footnote 23.

Table 1: Simulation Results from BLS Data

	First-Differences		Fixed Effects	
	AR(13)	AR(1)	AR(13)	AR(1)
CCE, G=N	0.1693	0.1590	0.2573	0.2528
CCE, G=4	0.0781	0.0714	0.0654	0.0657
CCE, G=2	0.0541	0.0582	0.0627	0.0645
Wild Bootstrap	0.1442	0.1553	0.2522	0.2459
Randomization Test	0.0950	0.0879	0.1306	0.1355
Perturbed Scores	0.1442	0.1332	0.2568	0.2513
CAI	0.0541	0.0493	0.0546	0.0483

Notes: Table reports rejection rates for 5% level tests from simulations using BLS unemployment data. All results are based on 10,000 simulations. CCE refers to the cluster covariance matrix estimator. A t -distribution with $G - 1$ degrees of freedom is used. Wild bootstrap uses each state as a cluster. Randomization Test refers to the Canay et al. (forthcoming) approach. Perturbed Scores is the unadjusted method discussed in Section 3.2. CAI refers to the correlation-adjusted score inference procedure introduced in this paper. Sample size is $N=49$, $T=383$.

improves the rejection rate to 0.054. The CAI procedure rejects at a rate of 0.054.

The next column presents the equivalent results for the AR(1) process. The rejection rates are generally similar for all the inference methods. CCE with $G = 2$ rejects at a rate of 0.058; the CAI rejection rate is 0.049. Next, I estimate equation (10) using state and time fixed effects. Most methods fare worse when fixed effects are used instead of first-differences. For the $G = 2$ case, the CCE rejection rates are 0.063 and 0.065. The CAI procedure has rejection rates of 0.055 and 0.048.

5.2 CGM Replication

Next, I replicate the simulated data used in Cameron et al. (2008). The data are generated in the following manner:

$$\begin{aligned}
 y_{st} &= \beta x_{st} + u_{st}, \\
 x_{st} &= z_s + z_{st}, \\
 u_{st} &= \epsilon_s + \epsilon_{st},
 \end{aligned}$$

where z_s , z_{st} , and ϵ_s are independent and distributed $N(0,1)$ and $\beta = 0$. I vary the distribution which generates ϵ_{st} . When the error distribution is labeled as normal, $\epsilon_{st} \sim$

$N\left(0, 9 \times (z_s + z_{st})^2\right)$. It is also instructive to show rejection rates when the error distribution is asymmetric so I also include simulations in which $\epsilon_{st} = 5 \times (z_s + z_{st}) \left[\mathcal{V} - \frac{1}{7}\right]$, where $\mathcal{V} \sim \text{Beta}(1, 6)$. I label the error distribution as Beta(1,6) in these cases.¹¹

I include simulations in which the number of states is equal to 6 and 10. I use even numbers so that the BCH method for $G = 2$ has an equal number of states in each cluster. Since there is no cross-sectional dependence, I randomize which states are included in each cluster when $G < N$. I set $T = 30$ in all simulations.

Table 2 presents the simulation results. In Columns (1)-(4), I estimate the specification using OLS regression. There is no cross-cluster dependence in these data so the independence assumption required by many of these methods holds. CCE with $G = 2$ has a rejection rate between 0.055 and 0.062, depending on the number of states and the distribution of the error term. The randomization test approach generates rejection rates between 0.049 and 0.061. The perturbed scores method introduced in this paper fares well with rates between 0.050 and 0.055. CAI also rejects at close to the expected rate, though it should be noted that the unadjusted version (perturbed scores) appears to perform slightly better, suggesting that imposing independence permits more accurate inference when that assumption is correct. However, it is worth highlighting that CAI rejects at close to the appropriate rate – between 0.053 and 0.057.

Using data generated as described above, I estimate the specification using median regression to illustrate the usefulness of the proposed inference procedure for nonlinear estimators. Parente and Santos Silva (2013) discusses quantile regression with clustered data. I use the standard errors generated by the `qreg2` command in Stata (Machado et al. (2011)) and adjust the t-statistics using BCH. The wild bootstrap method in Cameron et al. (2008) is only useful for linear models so I do not include wild bootstrap results for the median regression simulations.¹² The results are presented in columns (5) and (6). CCE overrejects slightly when $G \geq 4$. However, it underrejects in the $G = 2$ case. CAI rejects at close to the appropriate rate.

For Column (7)-(10), I generate data in the same manner but include state and time fixed effects when estimating the specification with OLS (even though these fixed effects are unnecessary for consistent estimation of β). The motivation for these simulations is the

¹¹This error term is asymmetric but mean zero.

¹²Hagemann (2016) introduces a wild bootstrap approach for quantile regression, but I do not include results from this method here.

Table 2: Simulation Results Replicating CGM

	OLS				Median Regression		OLS with Fixed Effects			
Number of Units	(1) N=6	(2) N=10	(3) N=6	(4) N=10	(5) N=6	(6) N=10	(7) N=6	(8) N=10	(9) N=6	(10) N=10
Error Distribution	Normal	Normal	Beta(1,6)	Beta(1,6)	Normal	Normal	Normal	Normal	Beta(1,6)	Beta(1,6)
CCE, G=N	0.0889	0.0885	0.1006	0.0890	0.0655	0.0661	0.0821	0.0685	0.0907	0.0773
CCE, G=4	0.0884	0.0710	0.0915	0.0738	0.0653	0.0588	0.0830	0.0636	0.0883	0.0639
CCE, G=2	0.0569	0.0551	0.0623	0.0555	0.0255	0.0320	0.0572	0.0541	0.0608	0.0550
Wild Bootstrap	0.0588	0.0546	0.0655	0.0556			0.0735	0.0516	0.0770	0.0562
Randomization Test	0.0546	0.0494	0.0612	0.0553	0.0593	0.0498	0.0062	0.0622	0.0083	0.0650
Perturbed Scores	0.0522	0.0531	0.0552	0.0500	0.0580	0.0525	0.0883	0.0661	0.0948	0.0728
CAI	0.0556	0.0542	0.0565	0.0527	0.0547	0.0504	0.0569	0.0504	0.0587	0.0554

Notes: Table reports rejection rates for 5% level tests from simulations. All results are based on 10,000 simulations with $T = 30$. CCE refers to the cluster covariance matrix estimator. A t -distribution with $G - 1$ degrees of freedom is used. Wild bootstrap uses each state as a cluster. Randomization Test refers to the Canay et al. (forthcoming) approach. Perturbed Scores is the unadjusted method discussed in Section 3.2. CAI refers to the correlation-adjusted score inference procedure introduced in this paper. When the error distribution is specified as Normal, $\epsilon_{st} \sim N\left(0, 9 \times (z_s + z_{st})^2\right)$. When the error distribution is specified as Beta, $\epsilon_{st} = 5 \times (z_s + z_{st}) \left[\mathcal{V} - \frac{1}{7}\right]$ with $\mathcal{V} \sim \text{Beta}(1, 6)$. When fixed effects are included in Columns (7)-(10), I form $\frac{N}{2}$ clusters with 2 units per cluster to implement the Randomization Test approach.

difficulty that fixed effects can cause for many inference procedures. For the randomization test method, I randomly choose two units for each cluster and estimate a fixed effects specification for that cluster. This approach cuts the number of clusters in half, which creates problems when the number of clusters is already small.

CCE with $G = 2$ has rejection rates between 0.054 and 0.061. The randomization test method struggles with power in the $N = 6$ case because there are too few possible t -statistics that can be simulated given $\frac{N}{2} = 3$ independent clusters.¹³ In the $N = 10$ case, this method appears to slightly overreject. CAI has close to the expected rejection rate - varying between 0.050 and 0.059. The correlation adjustment is important in the fixed effects specifications as the the unadjusted version of this approach, as one would expect, overrejects across all models.

5.3 Differences-in-Differences

The next set of simulations are motivated by those provided in Conley and Taber (2011) and also included in the Appendix of Canay et al. (forthcoming). Many inference methods struggle in the differences-in-differences context in which only a few units adopt the treatment. For these simulations, I provide results from the same methods as before, but I also include rejection rates generated by the method introduced in Conley and Taber (2011).¹⁴ For the randomization test method, I follow the Canay et al. (forthcoming) approach and estimate the parameter of interest separately for each treated state by estimating a fixed effects specification using that one treated state and all control states. For the perturbed score approach and CAI, it is only necessary to estimate the model once, but I take an equivalent approach and only use the scores for the treated states (the scores of the untreated states are uninformative). The data are generated in the following manner:

$$\begin{aligned} y_{it} &= \theta D_{it} + \beta Z_{jt} + \epsilon_{jt}, \\ \epsilon_{it} &= \rho \epsilon_{i,t-1} + \mu_{it}, \\ Z_{it} &= \gamma D_{it} + \nu_{it}, \end{aligned}$$

where $\theta = 0$, $\beta = 1$, $\gamma = 0.5$, and $\nu_{it} \sim N(0, 1)$ in all models. The null hypothesis is $H_0 : \theta = 0$. Let J_0 represent the set of untreated units and J_1 represent the set of

¹³This method produces p-value intervals which are large. I report the lower bound of these intervals.

¹⁴I use the $\hat{\Gamma}^*$ estimator (see page 117 of Conley and Taber (2011)).

treated units, where $|J_0| + |J_1| = 50$. The treatment variable is assigned in the following manner:

$$D_{it} = \begin{cases} 0 & \text{if } i \in J_0 \\ 0 & \text{if } i \in J_1 \text{ and } t < t_i^* \\ 1 & \text{if } i \in J_1 \text{ and } t \geq t_i^* \end{cases},$$

where $t_i^* = \min\{2i, T\}$. The remaining components of the data-generating process vary across the models. I implement seven different variations, each described below. Let $\mathcal{V} \sim \text{Beta}(1, 6)$,

- (a) $T = 10, |J_1| = 8, \rho = 0.5, \mu_{it} \sim N(0, 16)$
- (b) $T = 10, |J_1| = 8, \rho = 0.5, \mu_{it} \sim N(0, 1)$ if $i \leq 4, \mu_{it} \sim N(0, 16)$ if $i > 4$
- (c) $T = 10, |J_1| = 6, \rho = 0.5, \mu_{it} \sim N(0, 1)$ if $i \leq 4, \mu_{it} \sim N(0, 16)$ if $i > 4$
- (d) $T = 10, |J_1| = 8, \rho = 0.5, \mu_{it} = \mathcal{V} - \frac{1}{7}$ if $i \leq 4, \mu_{it} = 4 \times [\mathcal{V} - \frac{1}{7}]$ if $i > 4$
- (e) $T = 10, |J_1| = 8, \rho = 0.8, \mu_{it} \sim N(0, 1)$ if $i \leq 4, \mu_{it} \sim N(0, 16)$ if $i > 4$
- (f) $T = 5, |J_1| = 8, \rho = 0.5, \mu_{it} \sim N(0, 1)$ if $i \leq 4, \mu_{it} \sim N(0, 16)$ if $i > 4$
- (g) $T = 5, |J_1| = 8, \rho = 0.5, \mu_{it} = \mathcal{V} - \frac{1}{7}$ if $i \leq 4, \mu_{it} = 4 \times [\mathcal{V} - \frac{1}{7}]$ if $i > 4$

Table 3: Differences-in-Differences Simulation

Model	(a)	(b)	(c)	(d)	(e)	(f)	(g)
CCE, G=N	0.1116	0.1643	0.1765	0.1535	0.1674	0.0996	0.0970
CCE, G=4	0.0822	0.1051	0.1279	0.1033	0.1084	0.0714	0.0734
CCE, G=2	0.0601	0.0586	0.0714	0.0595	0.0612	0.0510	0.0545
Wild Bootstrap	0.0541	0.0589	0.0636	0.0562	0.0613	0.0486	0.0518
Randomization Test	0.0635	0.0611	0.0580	0.0554	0.0582	0.0551	0.0562
Conley-Taber	0.0724	0.3922	0.3825	0.3854	0.3957	0.3452	0.3505
Perturbed Scores	0.0485	0.0454	0.0467	0.0408	0.0383	0.1066	0.1033
CAI	0.0553	0.0519	0.0497	0.0527	0.0471	0.0502	0.0484

Notes: Table reports rejection rates for 5% level tests from simulations. All results are based on 10,000 simulations. CCE refers to the cluster covariance matrix estimator. A t -distribution with $G - 1$ degrees of freedom is used. Wild bootstrap uses each state as a cluster. Randomization Test refers to the Canay et al. (forthcoming) approach. Perturbed Scores is the unadjusted method discussed in Section 3.2. CAI refers to the correlation-adjusted score inference procedure introduced in this paper. Each model is described in the text. For Models (f) and (g), the CAI method creates clusters based on time and then adjusts for correlations within-units.

For most simulations, there are 8 treated states out of 50 states in total. The treated states adopt the policy at varying times during the sample period, setting up a straightforward differences-in-differences design. Table 3 presents the results of these simulations. Model (a) is most similar to the simulations shown in Conley and Taber (2011) as the errors across clusters are i.i.d. The Conley-Taber approach rejects at a rate of 0.072. CCE with $G = 2$ rejects at 0.060 while the wild bootstrap rejects at a rate of 0.054. The randomization test approach rejects at a slightly higher rate (0.064) while the perturbed scores approach rejects at close to the expected rate. CAI rejects at a rate of 0.055.

In Model (b), the error term distribution varies across units, which violates the assumption required by Conley and Taber (2011). Thus, the Conley-Taber method severely overrejects. The other methods have similar rejection rates as the rates for Model (a). CAI rejects at a rate of 0.052. Model (c) reduces the number of treated units to 6. The rejection rates for CCE appear to increase non-trivially, but the other methods are relatively unaffected. CAI rejects at a rate of 0.050.

Model (d) uses asymmetric errors. Rejection rates are relatively unaffected when compared to the symmetric errors in Model (b). CAI rejects at a rate of 0.053. Model (e) increases the within-cluster dependence of the error term. CAI rejects at a rate of 0.047. It is worth noting that the perturbed score approach without the correlation adjustment appears to under-reject in many of these simulations, but CAI rejects at appropriate levels.

Models (f) and (g) reduce the number of time periods to $T = 5$. For these models, the perturbed scores and CAI methods treat the time dimension as clusters. CAI adjusts for dependence within-units. This approach was discussed in Section 4.3.1. The CAI rejection rates are 0.050 when the error terms are normally-distributed and 0.048 when they are asymmetric.

5.4 Correlated Clusters

Clusters may be correlated for a variety of reasons that are unrelated to the inclusion of time fixed effects in the estimated specification. Here, I generate observations defined by state i in group g at time t . The motivation for these simulations is to create highly-correlated

Table 4: Correlated Units

	Model 1		Model 2			
	N=8, T=10	N=8, T=30	N=6, T=10	N=10, T=30	N=10, T=50	N=30, T=50
CCE, G=N	0.8161	0.7601	0.4963	0.3413	0.2692	0.2042
CCE, G=4	0.8079	0.7421	0.3781	0.2221	0.2060	0.1357
CCE, G=2	0.5895	0.5684	0.3682	0.1861	0.1927	0.0782
Wild Bootstrap	0.8262	0.7724	0.2706	0.2134	0.2135	0.1783
Randomization Test	0.5408	0.5337	0.2507	0.2004	0.2025	0.1763
Perturbed Scores	0.6846	0.6866	0.2599	0.2092	0.2082	0.1810
CAI	0.0510	0.0489	0.0521	0.0563	0.0506	0.0538

Notes: Table reports rejection rates for 5% level tests from simulations. All results are based on 10,000 simulations. CCE refers to the cluster covariance matrix estimator. A t -distribution with $G - 1$ degrees of freedom is used. Wild bootstrap uses each state as a cluster. Randomization Test refers to the Canay et al. (forthcoming) approach. Perturbed Scores is the unadjusted method discussed in Section 3.2. CAI refers to the correlation-adjusted score inference procedure introduced in this paper. In Model 1, there are 2 groups, each with 4 highly-dependent units. In Model 2, there are $\frac{N}{2}$ groups, each with 2 highly-dependent units.

states, but the researcher does not know which states are independent.

$$\begin{aligned}
y_{igt} &= \beta x_{igt} + u_{igt}, \\
x_{igt} &= z_{gt} + 0.1 \times \nu_{igt}, \\
z_{gt} &= .9 \times z_{g,t-1} + \phi_{gt}, \\
u_{igt} &= u_{gt} + 0.1 \times \zeta_{igt}, \\
u_{gt} &= .9 \times u_{g,t-1} + \eta_{gt},
\end{aligned}$$

where ϕ_{gt} , ν_{igt} , ζ_{igt} , and η_{gt} are independent and distributed $N(0, 1)$. I set $\beta = 0$. x_{igt} represents a simulated treatment variable that varies by state i in group g at time t . Both x and u are highly-correlated within each group; z and ω are, respectively, common components of these random variables. When creating the clusters, I randomly-assign states to clusters such that states within the same group may not be in the same cluster when estimating CCE. I expect this to cause difficulty for most inference approaches. The CAI method, however, estimates the relationship between the scores of each state such that it is unnecessary to decide which states should be in the same cluster. For “Model 1,” I generate the above data for $N = 8$ where there are 2 groups with 4 states in each group. “Model 2” includes two states per group, and I vary the number of groups. Note that even if the true clusters were known, these data generating processes would still create difficulties for many inference methods since there are few independent clusters. CAI, however, extracts the useful information from each state, generating N independent functions.

Given that correlated states are not correctly placed in the same cluster for CCE, it is not surprising that the BCH method severely overrejects, as shown in Table 4. For Model 1, the rejection rates are at least 56% using CCE. The other approaches also reject at especially high rates. The results are included to illustrate that CAI produces appropriate rates even when alternative methods reject more often than they do not. CAI rejects at close to the appropriate rate.

The rejection rates using CCE for Model 2 are smaller, though still very high in most cases. The other approaches severely overreject in all cases. CAI has rejection rates between 5.0% and 5.6% even in the presence of highly-correlated clusters.

5.5 Power Curves

Using the data-generating process described in Section 5.1, I study rejection rates when $\beta \neq 0$. I use the AR(13) process and include state fixed effects. I set β equal to a constant and generate a new outcome variable, adding βx_{st} to $\ln(y_{st})$. Figure 1 shows the rejection rate for each constant. I compare the power curve of CAI to the rejection rates using CCE with $G = 2$ since relative to the other (non-CAI) methods, CCE with $G = 2$ had the lowest rejection rate (0.063) in the $\beta = 0$ case.

Figure 1: Power Curve, BCH Data, AR(13) Process with State Fixed Effects

The rejection rate for CAI is always higher than the CCE rejection rate except for when the null hypothesis is true (as shown previously in Table 1). This pattern of results occurs for two reasons. First, CAI explicitly adjusts for correlations across clusters so it not surprising that it rejects at close to the expected rate when $\beta = 0$. Second, in the $\beta \neq 0$ case, CAI has more power than CCE because CCE is permitting arbitrary within-cluster dependence where many states compose each cluster. CAI performs inference at the state-level, not the levels of the clusters (as defined for the CCE, $G = 2$ approach), using independent variation from each state.

5.6 Disemployment Effects of the Minimum Wage

An influential literature studies whether increases in the minimum wage affect the employment rate of teenagers. I estimate a specification often found in this literature which models the log of the 16-19 year old employment rate as a linear function of state fixed effects, time fixed effects, and the log of the state minimum wage.¹⁵ While there is a debate about the appropriateness of this specification (see Allegretto et al. (forthcoming); Meer and West (2015); Neumark et al. (2014a,b)), I focus on inference given this model.¹⁶ The literature typically adjusts standard errors to account for within-state clustering. However, there are many reasons that it may be inappropriate to consider each state in the data as independent. For example, employment rates and minimum wage policy in states with similar industry compositions may respond to national economic shocks in correlated ways. It is also common in the literature to estimate specifications which include interactions based on Census region and time. While this approach may better account for common regional shocks, it also has the potential to mechanically induce more dependence across states within the same Census region.

Following the literature, I use employment rates from the Current Population Survey.¹⁷ I provide confidence intervals generated by CCE, defining clusters as states since this is typical in the literature. I also use CAI and invert the test statistics to create confidence intervals.

Table 5 presents the results. The estimated disemployment elasticity is -0.27 or -0.29, depending on whether time fixed effects or region-time interactions are used. These estimates are similar to the typical estimates found in the literature. Assuming that states are independent and using CCE to estimate the variance-covariance matrix, these estimates are statistically significant from zero. In the first model with time fixed effects, the 95% confidence interval using CAI includes 0.00, though this is the top of the interval. When region-time fixed effects are included, the confidence interval becomes especially large, suggesting that there is substantial dependence across observations in the same region. The inclusion

¹⁵The literature also typically includes controls for the state unemployment rate and the log of the size of the 16-19 population while weighting by population. To simplify the analysis here, I perform unweighted regressions and do not control for these additional covariates. These changes have little effect on the estimates.

¹⁶Concerns about selecting appropriate control groups for each state are addressed in complementary work (Powell (2016)).

¹⁷Code and data can be found at Arindrajit Dube's site: <https://arindube.com/working-papers/> (last accessed January 15, 2016)

Table 5: Disemployment Effects of the Minimum Wage

Outcome:	ln(Teen employment rate)	
Model:	Time fixed effects	Region-time fixed effects
Coefficient Estimate	-0.27	-0.29
CCE, G=N	[-0.43, -0.11]	[-0.47, -0.11]
CAI	[-0.53, 0.00]	[-0.40, 0.51]

Notes: N=7056. 95% confidence intervals are presented in brackets. State fixed effects also included. Coefficient estimate is the parameter associated with the log of the minimum wage.

of region-time fixed effects eliminates important sources of variation and reduces power. While it has become common in the minimum wage literature to account for geographic-time shocks, this result suggests that these estimates are not informative as the confidence intervals cannot rule out large negative or positive effects.

6 Discussion and Conclusion

This paper introduces a simple method to conduct inference with correlated, heterogeneous clusters which is valid even when the number of clusters is small. Clusters may be correlated for several reasons including economic (due to spatial proximity or similar industry compositions) or econometric (due to the inclusion of time fixed effects). The procedure uses a Wald test statistic and simulates the distribution of this statistic under the null hypothesis. This simulation involves perturbing the scores, without re-estimating the model, and is computationally fast to implement. This approach has several advantages over traditional inference techniques and allows for different-sized (with individual-level data) and heterogeneous clusters without requiring estimation of the full model separately for each cluster.

The main advantage of CAI is that it allows for strong dependence across clusters. Given arbitrary within-unit dependence and correlations across clusters, it is possible that all observations are dependent and typical inference methods are inappropriate. CAI uses the panel nature of the data to estimate the relationships between the clusters using OLS. Once these relationships are estimated, independent functions can be isolated and inference can proceed as if the clusters were uncorrelated. This method should be useful generally for linear and nonlinear estimators with panel data. This procedure improves upon approaches

which require parameterizing or at least knowledge of the source of dependence (e.g., spatial). CAI does not require prior knowledge about which clusters may be dependent. Simulations illustrate that CAI rejects at appropriate rates even in the presence of highly-dependent clusters in which other methods reject at rates greater than 50%.

A Appendix

Consistency: Assume that **A1-A2** hold, then $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.

Let $\hat{\mathcal{L}}(\boldsymbol{\theta}) = \hat{Q}(\boldsymbol{\theta}) + a(\boldsymbol{\theta})'\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ is an $r \times 1$ vector of Lagrange multipliers. The constrained estimator maximizes this equation. Note that all the conditions in Theorem 2.1 of Newey and McFadden (1994) hold using $\hat{\mathcal{L}}(\boldsymbol{\theta})$ in place of $\hat{Q}(\boldsymbol{\theta})$ given **A1** and **A2**.

Lemma 3.1. Assume that **A1-A6** hold with $T \rightarrow \infty$. Then,

$$\begin{aligned} \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \nabla_{\boldsymbol{\theta}} Q_{it}(\tilde{\boldsymbol{\theta}}) &\xrightarrow{d} N(0, \tilde{V}), \\ \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T W_i \nabla_{\boldsymbol{\theta}} Q_{it}(\tilde{\boldsymbol{\theta}}) &\xrightarrow{d} N(0, \tilde{V}), \end{aligned}$$

where $\tilde{V} = (AJ^{-1}A')^{-1}AJ^{-1}VJ^{-1}A'(AJ^{-1}A')^{-1}$.

Proof. As before, let

$$\hat{\mathcal{L}}(\boldsymbol{\theta}) = \hat{Q}(\boldsymbol{\theta}) + a(\boldsymbol{\theta})'\boldsymbol{\gamma},$$

where $\boldsymbol{\gamma}$ is an $r \times 1$ vector of Lagrange multipliers. The constrained estimator maximizes the above equation, solving the following set of equations:

$$\begin{pmatrix} \sqrt{NT} \nabla_{\boldsymbol{\theta}} \hat{Q}(\tilde{\boldsymbol{\theta}}) + \nabla_{\boldsymbol{\theta}} a(\tilde{\boldsymbol{\theta}})' \sqrt{NT} \tilde{\boldsymbol{\gamma}} \\ \sqrt{NT} a(\tilde{\boldsymbol{\theta}}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

A mean value expansion around $\boldsymbol{\theta}_0$ gives

$$\sqrt{NT} \nabla_{\boldsymbol{\theta}} \hat{Q}(\tilde{\boldsymbol{\theta}}) = \sqrt{NT} \nabla_{\boldsymbol{\theta}} \hat{Q}(\boldsymbol{\theta}_0) + \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \hat{Q}(\boldsymbol{\theta}^*) \sqrt{NT} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \quad (11)$$

$$\sqrt{NT} a(\tilde{\boldsymbol{\theta}}) = \sqrt{NT} \nabla_{\boldsymbol{\theta}} a(\boldsymbol{\theta}^{**}) (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \quad (12)$$

where $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^{**}$ represent vectors of points between $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Then, by consistency of $\tilde{\boldsymbol{\theta}}$ and **A5**,

$$\begin{bmatrix} \sqrt{NT} \nabla_{\boldsymbol{\theta}} \widehat{Q}(\boldsymbol{\theta}_0) \\ 0 \end{bmatrix} + \begin{bmatrix} J & A' \\ A & 0 \end{bmatrix} \begin{bmatrix} \sqrt{NT} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ \sqrt{NT} \tilde{\boldsymbol{\gamma}} \end{bmatrix} + o_p = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Rearranging terms,

$$\begin{bmatrix} \sqrt{NT} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ \sqrt{NT} \tilde{\boldsymbol{\gamma}} \end{bmatrix} = - \begin{bmatrix} J^{-1} - J^{-1} A' (A J^{-1} A')^{-1} A J^{-1} & J^{-1} A' (A J^{-1} A')^{-1} \\ (A J^{-1} A')^{-1} A J^{-1} & -(A J^{-1} A')^{-1} \end{bmatrix} \begin{bmatrix} \sqrt{NT} \nabla_{\boldsymbol{\theta}} \widehat{Q}(\boldsymbol{\theta}_0) \\ 0 \end{bmatrix} + o_p.$$

From equation (11),

$$\begin{aligned} \sqrt{NT} \nabla_{\boldsymbol{\theta}} \widehat{Q}(\tilde{\boldsymbol{\theta}}) &= \sqrt{NT} \nabla_{\boldsymbol{\theta}} \widehat{Q}(\boldsymbol{\theta}_0) + J \sqrt{NT} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p \\ &= \sqrt{NT} \nabla_{\boldsymbol{\theta}} \widehat{Q}(\boldsymbol{\theta}_0) + J [-J^{-1} + J^{-1} A' (A J^{-1} A')^{-1} A J^{-1}] \sqrt{NT} \nabla_{\boldsymbol{\theta}} \widehat{Q}(\boldsymbol{\theta}_0) + o_p \\ &= (A J^{-1} A')^{-1} A J^{-1} \sqrt{NT} \nabla_{\boldsymbol{\theta}} \widehat{Q}(\boldsymbol{\theta}_0) + o_p. \end{aligned}$$

From **A3** and **A4**, it follows that $\frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \nabla_{\boldsymbol{\theta}} Q_{it}(\boldsymbol{\theta}_0) \xrightarrow{d} N(0, V)$. Therefore,

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \nabla_{\boldsymbol{\theta}} Q_{it}(\tilde{\boldsymbol{\theta}}) \xrightarrow{d} N(0, \tilde{V}),$$

where $\tilde{V} = (A J^{-1} A')^{-1} A J^{-1} V J^{-1} A' (A J^{-1} A')^{-1}$.

Given the conditions governing the weights and **A4**, $\nabla_{\boldsymbol{\theta}} Q_i(\boldsymbol{\theta}_0)$ and $W_i \nabla_{\boldsymbol{\theta}} Q_i(\boldsymbol{\theta}_0)$ have the same asymptotic variance:

$$\text{Var}(W_i \nabla_{\boldsymbol{\theta}} Q_i(\boldsymbol{\theta}_0)) = \text{Var}(W_i) \text{Var}(\nabla_{\boldsymbol{\theta}} Q_i(\boldsymbol{\theta}_0)) = \text{Var}(\nabla_{\boldsymbol{\theta}} Q_i(\boldsymbol{\theta}_0)).$$

By **A3(ii)** and **A6(iii)**,

$$E |W_i \nabla_{\boldsymbol{\theta}} Q_i(\boldsymbol{\theta}_0)|^r \leq \left(E |W_i|^{2r} \right)^{\frac{1}{2}} \left(E |\nabla_{\boldsymbol{\theta}} Q_i(\boldsymbol{\theta}_0)|^{2r} \right)^{\frac{1}{2}} < \Delta < \infty.$$

Thus, $\frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T W_i \nabla_{\boldsymbol{\theta}} Q_{it}(\boldsymbol{\theta}_0) \xrightarrow{d} N(0, V)$.

Following the same steps as above:

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T W_i \nabla_{\boldsymbol{\theta}} Q_{it}(\tilde{\boldsymbol{\theta}}) \xrightarrow{d} N(0, \tilde{V}),$$

□

Theorem 3.1. *Assume **A1-A6** hold with $T \rightarrow \infty$. Then, S and $S^{(k)}$ converge to the same distribution.*

Given the result in Lemma 3.1, $\frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T g_{it}(\tilde{\boldsymbol{\theta}})$ and $\frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T W_i g_{it}(\tilde{\boldsymbol{\theta}})$ converge to the same distribution. The result follows immediately from the continuous mapping theorem.

Theorem 4.1. *Assume **A1-A3**, **A7**, (i) $E|g_{ith}(\boldsymbol{\theta})|^2|^{(r+\delta)} < \Delta < \infty$ for all i, t, h and for some $\delta > 0$, for $\boldsymbol{\theta}$ in some neighborhood of $\boldsymbol{\theta}_0$; (ii) $g_{ith}(\boldsymbol{\theta})$ continuous at $\boldsymbol{\theta}_0$ with probability one for all i, t, h ; (iii) $E \left[\frac{1}{T} \sum_{t=1}^T g_{th}^i(\boldsymbol{\theta}_0) g_{th}^i(\boldsymbol{\theta}_0)' \right]$ is uniformly positive definite. Then,*

$$\text{Cov} \left[\frac{1}{T} \sum_{t=1}^T \left(g_{ith}(\tilde{\boldsymbol{\theta}}) - g_{th}^i(\tilde{\boldsymbol{\theta}})' \hat{\mathbf{b}}_h^i \right), \quad \frac{1}{T} \sum_{t=1}^T \left(g_{jth}(\tilde{\boldsymbol{\theta}}) - g_{th}^j(\tilde{\boldsymbol{\theta}})' \hat{\mathbf{b}}_h^j \right) \right] \xrightarrow{p} 0 \quad (13)$$

for all h and for all $i \neq j$.

Proof. By (i) and **A3**, the law of large numbers implies

$$\frac{1}{T} \sum_{t=1}^T g_{th}^i(\boldsymbol{\theta}_0) g_{th}^i(\boldsymbol{\theta}_0)' \xrightarrow{p} E \left[\frac{1}{T} \sum_{t=1}^T g_{th}^i(\boldsymbol{\theta}_0) g_{th}^i(\boldsymbol{\theta}_0)' \right]. \quad (14)$$

Moreover, given **A1** and **A2**, $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$. This result, equation (14), and conditions (i) and (ii) imply

$$\frac{1}{T} \sum_{t=1}^T g_{th}^i(\tilde{\boldsymbol{\theta}}) g_{th}^i(\tilde{\boldsymbol{\theta}})' \xrightarrow{p} E \left[\frac{1}{T} \sum_{t=1}^T g_{th}^i(\boldsymbol{\theta}_0) g_{th}^i(\boldsymbol{\theta}_0)' \right].$$

This result holds by Lemma 4.3 in Newey and McFadden (1994). Similarly,

$$\frac{1}{T} \sum_{t=1}^T g_{th}^i(\tilde{\theta}) g_{ith}(\tilde{\theta}) \xrightarrow{p} E \left[\frac{1}{T} \sum_{t=1}^T g_{th}^i(\theta_0) g_{ith}(\theta_0) \right]$$

By assumption (iii),

$$\begin{aligned} & \left[\frac{1}{T} \sum_{t=1}^T g_{th}^i(\tilde{\theta}) g_{th}^i(\tilde{\theta})' \right]^{-1} \left[\frac{1}{T} \sum_{t=1}^T g_{th}^i(\tilde{\theta}) g_{ith}(\tilde{\theta}) \right] \\ & \xrightarrow{p} \mathbf{b}_h^i + E \left[\frac{1}{T} \sum_{t=1}^T g_{th}^i(\theta_0) g_{th}^i(\theta_0)' \right]^{-1} E \left[\frac{1}{T} \sum_{t=1}^T g_{th}^i(\theta_0) \mu_{ith}(\theta_0) \right] = \mathbf{b}_h^i \end{aligned}$$

Thus, $\hat{\mathbf{b}}_h^i \xrightarrow{p} \mathbf{b}_h^i$. Then, it is straightforward to show that $\hat{\mu}_{ith} \xrightarrow{p} \mu_{ith}$ and the result follows by **A7**:

$$\text{Cov} \left(\frac{1}{T} \sum_{s=1}^T \hat{\mu}_{ish}, \frac{1}{T} \sum_{s=1}^T \hat{\mu}_{jsh} \right) \xrightarrow{p} E \left[\left(\frac{1}{T} \sum_{s=1}^T \mu_{ish} \right) \left(\frac{1}{T} \sum_{s=1}^T \mu_{jsh} \right) \right] = 0.$$

□

References

- Aitchison, John and SD Silvey, “Maximum-likelihood estimation of parameters subject to restraints,” *The Annals of Mathematical Statistics*, 1958, pp. 813–828.
- Allegretto, Sylvia A, Arindrajit Dube, Michael Reich, and Ben Zipperer, “Credible research designs for minimum wage studies,” *Industrial and Labor Relations Review*, forthcoming.
- Arellano, M, “Computing Robust Standard Errors for Within-Groups Estimators,” *Oxford Bulletin of Economics and Statistics*, 1987, 49 (4), 431–34.
- Bakirov, Nail Kutluzhanovich and GJ Székely, “Students t-test for Gaussian scale mixtures,” *Journal of Mathematical Sciences*, 2006, 139 (3), 6497–6505.
- Barrios, Thomas, Rebecca Diamond, Guido W Imbens, and Michal Kolesar, “Clustering, spatial correlations, and randomization inference,” *Journal of the American Statistical Association*, 2012, 107 (498), 578–591.
- Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, “How Much Should We Trust Differences-in-Differences Estimates?,” *The Quarterly Journal of Economics*, 2004, 119 (1), 249–275.
- Bester, C Alan, Timothy G Conley, and Christian B Hansen, “Inference with dependent data using cluster covariance estimators,” *Journal of Econometrics*, 2011, 165 (2), 137–151.
- , –, –, Timothy J Vogelsang et al., “Fixed-b asymptotics for spatially dependent robust nonparametric covariance matrix estimators,” *Econometric Theory*, 2016, 32 (01), 154–186.
- Cameron, A Colin, Jonah B Gelbach, and Douglas L Miller, “Bootstrap-based improvements for inference with clustered errors,” *The Review of Economics and Statistics*, 2008, 90 (3), 414–427.
- Canay, Ivan A, Joseph P Romano, and Azeem M Shaikh, “Randomization Tests under an Approximate Symmetry Assumption,” *Econometrica*, forthcoming.
- Carter, Andrew V, Kevin T Schnepel, and Douglas G Steigerwald, “Asymptotic behavior of a t test robust to cluster heterogeneity,” *The Review of Economics and Statistics*, forthcoming.
- Conley, Timothy G, “GMM estimation with cross sectional dependence,” *Journal of Econometrics*, 1999, 92 (1), 1–45.
- and Christopher R Taber, “Inference with “difference in differences” with a small number of policy changes,” *The Review of Economics and Statistics*, 2011, 93 (1), 113–125.

- Donald, Stephen G and Kevin Lang**, “Inference with difference-in-differences and other panel data,” *The Review of Economics and Statistics*, 2007, *89* (2), 221–233.
- Driscoll, John C and Aart C Kraay**, “Consistent covariance matrix estimation with spatially dependent panel data,” *Review of economics and statistics*, 1998, *80* (4), 549–560.
- Hagemann, Andreas**, “Cluster-robust bootstrap inference in quantile regression models,” *Journal of the American Statistical Association*, 2016, (just-accepted), 1–30.
- Hansen, Christian B**, “Asymptotic properties of a robust variance matrix estimator for panel data when T is large,” *Journal of Econometrics*, 2007, *141* (2), 597–620.
- Ibragimov, Rustam and Ulrich K Müller**, “t-Statistic based correlation and heterogeneity robust inference,” *Journal of Business & Economic Statistics*, 2010, *28* (4), 453–468.
- **and –**, “Inference with few heterogeneous clusters,” *Review of Economics and Statistics*, 2016, *98* (1), 83–96.
- Kiefer, Nicholas M and Timothy J Vogelsang**, “Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation,” *Econometrica*, 2002, pp. 2093–2095.
- **and –**, “A new asymptotic theory for heteroskedasticity-autocorrelation robust tests,” *Econometric Theory*, 2005, *21* (6), 1130.
- Kim, Min Seong and Yixiao Sun**, “Heteroskedasticity and spatiotemporal dependence robust inference for linear panel models with fixed effects,” *Journal of Econometrics*, 2013, *177* (1), 85–108.
- Kline, Patrick and Andres Santos**, “A score based approach to wild bootstrap inference,” *Journal of Econometric Methods*, 2012, *1* (1), 23–41.
- Koenker, Roger W and Gilbert Bassett**, “Regression Quantiles,” *Econometrica*, January 1978, *46* (1), 33–50.
- Liang, Kung-Yee and Scott L Zeger**, “Longitudinal data analysis using generalized linear models,” *Biometrika*, 1986, pp. 13–22.
- Machado, J.A.F., P.M.D.C Parente, and J.M.C. Santos Silva**, “QREG2: Stata module to perform quantile regression with robust and clustered standard errors,” Statistical Software Components, Boston College Department of Economics November 2011.
- MacKinnon, James G and Matthew D Webb**, “Wild bootstrap inference for wildly different cluster sizes,” *Journal of Applied Econometrics*, 2016.
- Meer, Jonathan and Jeremy West**, “Effects of the minimum wage on employment dynamics,” *Journal of Human Resources*, 2015.

- Neumark, David, JM Ian Salas, and William Wascher**, “More on recent evidence on the effects of minimum wages in the United States,” *IZA Journal of Labor Policy*, 2014, 3 (1), 24.
- , – , **and** – , “Revisiting the Minimum Wage–Employment Debate: Throwing Out the Baby with the Bathwater?,” *Industrial & Labor Relations Review*, 2014, 67 (3 suppl), 608–648.
- Newey, Whitney K. and Daniel McFadden**, “Large sample estimation and hypothesis testing,” *Handbook of Econometrics*, 1994, 4, 2111–2245.
- Newey, Whitney K and Kenneth D West**, “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix,” *Econometrica*, 1987, 55 (3), 703–708.
- Parente, Paulo MDC and João Santos Silva**, “Quantile regression with clustered data,” *Journal of Econometric Methods*, 2013.
- Powell, David**, “Synthetic Control Estimation Beyond Case Studies: Does the Minimum Wage Reduce Employment?,” 2016.
- Rao, C Radhakrishna**, “Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation,” in “Mathematical Proceedings of the Cambridge Philosophical Society,” Vol. 44 Cambridge Univ Press 1948, pp. 50–57.
- Vogelsang, Timothy J**, “Heteroskedasticity, autocorrelation, and spatial correlation robust inference in linear panel models with fixed-effects,” *Journal of Econometrics*, 2012, 166 (2), 303–319.
- Webb, Matthew D**, “Reworking wild bootstrap based inference for clustered errors,” Technical Report, Queen’s Economics Department Working Paper 2014.
- White, Halbert**, *Asymptotic Theory for Econometricians*, Academic press, 2001.
- Wooldridge, Jeffrey M**, “Cluster-sample methods in applied econometrics,” *American Economic Review*, 2003, pp. 133–138.