Inference with Correlated Clusters

David Powell*

December 2015

Abstract

This paper introduces a method which permits valid inference given a finite number of heterogeneous, correlated clusters. It is common in empirical analysis to use inference methods which assume that each unit is independent. Panel data permit this assumption to be relaxed as it is possible to estimate the correlations across clusters and isolate the independent variation in each cluster for proper inference. Clusters may be correlated for many reasons such as geographic proximity, similar institutions, comparable industry compositions, etc. Moreover, with panel data, it is typical to include time fixed effects, which mechanically induce correlations across clusters. The introduced inference procedure uses a Wald statistic and simulates the distribution of this statistic in a manner that is valid even for a small number of clusters. To account for correlations across clusters, the relationship between each cluster is estimated and only the independent component of each cluster is used. The method is simple to use and only requires one estimation of the model. It can be employed for linear and nonlinear estimators. I present several sets of simulations and show that the inference procedure consistently rejects at the appropriate rate, even in the presence of highly-correlated clusters in which traditional inference methods severely overreject.

Keywords: Finite Inference, Correlated Clusters, Fixed Effects, Panel Data, Wald Statistic

JEL classification: C12, C15, C21, C22, C23

*RAND, dpowell@rand.org
1 Introduction

This paper introduces an inference procedure that is valid for a fixed number of heterogeneous, correlated clusters. The procedure is useful for panel data applications with arbitrary within-cluster dependence in which the clusters themselves are dependent. Many empirical applications rely on panel data and assume that units are independent. This assumption can be relaxed. There are many reasons that we may not want to assume that clusters are independent. 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 can be used for linear and nonlinear estimation and is valid for small $N$.

Clusters may be correlated for several reasons such as spatial proximity, similar political institutions, and comparable industry compositions. One motivation for the approach in this paper is the popularity of panel data applications which include unit (e.g., state) and time fixed effects. The inclusion of time fixed effects in linear models is equivalent to demeaning the outcome and explanatory variables. This demeaning creates dependence across all units within the same time period. Given arbitrary correlations within each unit and dependence across units, it is possible that the correlation between all observations is non-zero. With panel data, it is unnecessary to pre-determine which clusters are correlated or model the source of dependence. Instead, this dependence can be measured across all clusters 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 [1987b] introduced a “plug-in” covariance matrix for dependence over time. Conley [1999] discussed an equivalent procedure for spatially-dependent data. It has become common in applied work to use a cluster covariance estimator (CCE) which permits arbitrary within-cluster dependence (Liang and Zeger [1986], Arellano et al. [1987]). The importance of this adjustment was popularized in Bertrand et al. [2004] while the limitations of inference with only a few independent clusters is considered further in Donald and Lang [2007] and Wooldridge [2003]. As discussed in MacKinnon and Webb [2014], general results on CCE estimation assume that the number of clusters approaches infinity, the within-cluster error correlations are constant across clusters, and each cluster is the same size. The inference procedure introduced in this paper will not require any of these assumptions.
Recent work has considered the limitations of CCE while others have formalized assumptions under which CCE is valid even when the number of clusters is small. Hansen [2007] provides conditions under which CCE provides valid inference for fixed $N$ as $T \to \infty$. 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. [2013] considers the effects of heterogeneity (in size and other factors) on inference, finding that clusters of very different sizes can affect inference in finite samples. Similarly, Conley and Taber [2011] points out that traditional inference methods are not valid in the context of linear difference-in-differences models in which only a few groups adopt the policy of interest. They introduce a valid inference procedure for this context under the assumption that the errors across groups are identically and independently distributed.

Ibragimov and Müller [2010] introduces an inference procedure that 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 parameters separately for each 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%. This procedure is also only appropriate for tests involving a scalar, ruling out joint hypothesis tests. Finally, the procedure requires estimating the model separately for each cluster. This requirement disallows models which require variation across clusters to estimate some of the parameters, such as year fixed effects in many panel data applications. Ibragimov and Müller [2015] extends the test to two-sample problems and discusses applications to difference-in-differences estimation, but the general requirement of estimating parameters separately across clusters remains. Canay et al. [2014] builds on this approach, developing randomization tests which also require estimating parameters separately for each cluster. 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, ruling out time fixed effects and estimating common effects for parameters across clusters.

While most inference procedures require independence across clusters, 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].
With dependence across observations that are “close” in space and/or time, BCH considers conditions under which the clusters are asymptotically independent. As the size of the clusters grows, most observations in clusters are far from each other on these dimensions. However, generating test statistics using CCE under these conditions requires additional assumptions of homogeneity across clusters of both \(x_g'x_g\) and \(x_g'\epsilon_g\) in the linear case (similar assumptions are required in Hansen [2007]). These assumptions rule out typical difference-in-differences empirical strategies and many other cases where heterogeneity in these matrices is likely common. Other assumptions, such as equal-sized clusters, are also enforced. Furthermore, fixed effects potentially pose problems in this set-up (see footnote 21 of the paper). For panel data analysis, BCH recommends creating large spatially-grouped clusters. As an example, they consider a simulated policy which is spatially-correlated across states. They use CCE but suggest creating larger clusters, dividing the United States into two clusters\(^1\) and using a \(t\)-distribution with one degree of freedom. Simulations in BCH suggest that this method outperforms the Kiefer and Vogelsang [2002, 2005] approach.

Other work has also considered dependence on multiple dimensions. Cameron et al. [2011] considers multi-way clustering, recognizing that correlations may occur both spatially and temporally. The resulting estimator requires that each dimension is large and assumes independence across observations that do not belong to any common clusters. In contrast, this paper introduces a method which allows for dependence across (potentially) all observations, even when the number of clusters is small.

The literature has also proposed and analyzed bootstrap methods when the number of clusters is small. Cameron et al. [2008] presents simulations that show that the wild bootstrap method which simulates the distribution of a \(t\)-statistic and imposes the null hypothesis works well even for a few clusters. This method is robust to heterogeneity across clusters. The proposed inference method of this paper will have some similarities with the wild bootstrap method as it requires perturbing cluster-specific functions. However, this paper’s approach has computational advantages over bootstrap methods because it does not require re-estimating the model several times. Simulations provide evidence that it outperforms traditional inference methods, including the wild bootstrap method.

This paper develops a correlation-adjusted inference (CAI) procedure which uses the scores of M-estimators to derive valid p-values. I use a Wald Statistic and simulate the dis-
\(^1\)They test the use of CCE with other sizes as well.
tribution of this Wald Statistic under the null hypothesis. With asymptotically independent clusters, this simulated distribution can be used to generate valid p-values. However, I consider the possibility that the clusters are dependent. There are many possible reasons that clusters may be dependent. Time fixed effects can generate mechanical correlations across clusters. Alternatively, clusters may be correlated for economic reasons such as common shocks that affects states that are geographically close or have similar industry compositions. The inference procedure uses the panel nature of the data to estimate the relationship between the clusters and does not require knowledge of which clusters may be correlated, permitting correlations across all clusters. Given these estimates, it is possible to create \( N \) asymptotically (as \( T \to \infty \)) independent clusters using only the independent parts of the scores for each cluster. Given asymptotically uncorrelated clusters, use of the Wald Statistic and its simulated counterparts is appropriate.

The score-based inference procedure is also computationally attractive and has many advantages even in the case where clusters are uncorrelated. First, it simulates the test statistic using the scores derived from one estimation of the model. Relative to bootstrap techniques, this approach may be much less computationally-intensive. Second, using only the scores circumvents estimation of the Hessian matrix, which can be difficult for some nonlinear estimators. For example, inference using quantile regression depends on the reciprocal of the conditional density of the outcome variable and it is common to use kernel estimation methods. These methods depend on bandwidth choices and may be difficult to implement. The inference procedure introduced in this paper avoids the need for these bandwidth choices.

Kline and Santos [2012] introduces a score-based bootstrap which simulates the distribution of the parameters by randomly weighting the score functions. This procedure has computational advantages and is most similar in spirit to wild bootstrap methods with the added benefit that the entire model does not need to be estimated each time to produce simulated values of the parameters. The perturbed scores are multiplied by the inverse Hessian to generate new estimates. The procedure is not developed to account for dependence.

This paper makes several contributions. First, it develops an inference procedure which is straightforward to implement for linear and nonlinear models and a fixed number of clusters. For many extremum estimators, the procedure is simpler to implement than traditional estimation of the variance-covariance matrix because it only requires calculation of the score function. Using only the scores can reduce the required assumptions to generate
valid p-values. For example, many finite \( N \) inference methods require additional assumptions such as homogeneity (across clusters) of the Hessian and homogeneity of the scores. These restrictions will not be necessary in this paper. The clusters can be heterogeneous and different sizes. Second, the inference procedure permits valid inference in the presence of correlated clusters and small \( N \) (or small \( T \)). 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 \textit{a priori} knowledge about which clusters are possibly correlated but, instead, empirically estimates the correlations across all clusters. In one set of simulations, I generate two units per unknown cluster where the units are highly-dependent. The procedure provides appropriate inference even in this case. This procedure is simple to implement as estimating the dependence across clusters only requires OLS estimation.

The method uses constrained maximization of the objective function under the null hypothesis. Assuming a central limit theorem applies, the sum of the scores for each cluster divided by \( \sqrt{T} \) converges to a normal distribution with mean zero. The test statistic is a Wald Statistic using the mean of the scores across clusters normalized by the variance of the scores. The next step is to generate weights which are mean zero with unit variance, independent of the data, and equal within each cluster. If the clusters are independent, then the weighted average of scores normalized by the weighted variance has the same distribution as the original test statistic. Consequently, we can simulate the distribution of the test statistic. Large values of the test statistic relative to the simulated distribution imply that the null hypothesis can be rejected.

This paper is motivated by the possibility that clusters are correlated and I consider two general reasons that such correlations may occur. First, clusters may experience similar shocks over time that are not appropriately modeled by time fixed effects. Conley [1999] models spatial dependence explicitly while BCH discusses proper inference when such dependence converges to zero such that spatial clusters are asymptotically independent. These methods typically assume that dependence is stronger for units that are close together. Barrios et al. [2012] discusses conditions under which spatial correlations must be considered or can be ignored, again modeling dependence as a function of distance. With panel data, this assumption is unnecessary. It is possible to observe the relationship between the scores for each cluster as they covary over time. Once this relationship is estimated, the independent components of each cluster can be isolated for inference. This allows for adjustments for dependence across clusters regardless of the spatial proximity of the clusters. Clusters may
be correlated due to geographic proximity, similar industry compositions, comparable political institutions, etc. The method proposed in this paper does not require any knowledge or assumptions about the possible mechanisms driving co-movements between clusters and, instead, takes an empirical approach to determine these correlations.

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 when such time fixed effects are excluded. Including time fixed effects in linear models is equivalent to transforming the residual:

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

This transformation generates correlations across clusters since, typically,

$$\text{Corr} \left( \epsilon_{it} - \frac{1}{N} \sum_{j=1}^{N} \epsilon_{jt}, \epsilon_{kt} - \frac{1}{N} \sum_{j=1}^{N} \epsilon_{jt} \right) \neq 0.$$

As discussed in Hansen [2007], this demeaning creates a correlation across clusters of order $O(\frac{1}{N})$. This dependence disappear as $N \to \infty$ but many empirical applications involve a fixed number of units. Cameron and Miller [2015] discusses the importance of setting a large fractions of error correlations to 0 by assumption for proper variance estimation. With clustered errors, observations in different clusters are assumed to have zero correlation. With time fixed effects and arbitrary within-cluster error dependence, it is possible that there are no observations with independent errors. In this paper, I discuss an inference procedure that is valid even in this case by determining the correlations across clusters and isolating the independent variation in each cluster.

In the next section, I discuss the general framework and introduce the inference procedure for 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 3 considers correlated clusters and extends the method by creating independent functions from correlated clusters. The motivation for accounting for correlations across clusters should be clear and the benefits of estimating these correlations extends beyond the CAI method. The correlation
adjustment in Section 3 is appropriate for other inference procedures as well. Section 4 includes several simulations and compares CAI to other inference methods in the literature. I conclude in Section 5.

2 Model

This paper considers extremum estimators of the form

$$\theta = \arg\max_{\theta} Q(\theta),$$

(2)

where $Q(\theta)$ represents the objective function. The estimated parameters maximize the sample equivalent:

$$\hat{\theta} = \arg\max_{\theta} \hat{Q}(\theta),$$

(3)

with $\hat{Q}(\theta)$ representing the sample objective function (suppressing the dependence of this function on the data). This paper performs a hypothesis test which only requires estimation of $\theta$ under the null hypothesis, which is represented by

$$H_0 : a(\theta) = 0.$$  

(4)

This notation follows Newey and West [1987a] and allows for nonlinear hypotheses. The method proposed in this paper uses only the restricted estimator. In Section 3.3.1, I discuss a similar method which uses both restricted and unrestricted estimation of $\theta$ and is asymptotically equivalent to the method discussed here. In simulations, both methods work well. The restricted estimator is defined by

$$\tilde{\theta} = \arg\max_{\theta} \tilde{Q}(\theta) \quad \text{subject to} \quad a(\theta) = 0.$$  

(5)

Furthermore, I assume that there exists a vector that converges asymptotically to a mean-zero normal distribution with finite variance under the null hypothesis. Each element of this vector must be non-zero when the null hypothesis is not true. I label this function as $g(\theta)$. This function will be elements of the score function or gradient ($\nabla_{\theta}Q(\theta)$) in many cases, such as with maximum likelihood estimators. For GMM, a moment condition itself suffices. I refer to these functions as “score functions” throughout this paper, though other functions will meet the formal assumptions that are required for the inference procedure.
2.1 Examples

I give examples of three estimators and the corresponding functions that can be used for inference. First, ordinary least squares (OLS) chooses the parameters that minimize the sum of the square of the residuals. Second, Instrumental Variables Quantile Regression (IVQR) can be implemented using GMM. Third, probit regression is typically implemented through maximum likelihood.

Consider estimation of specification

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

under the null hypothesis \( \beta_1 = 0 \). Imposing the null hypothesis is equivalent to OLS estimation of

\[ y_{it} = \alpha + \beta_2 x_{it}^{(2)} + \epsilon_{it}. \]

Let \( \mathbf{x}_{it}' \tilde{\beta} = \tilde{\alpha} + \tilde{\beta}_2 x_{it}^{(2)} \). The score function for unit \( i \) under the null hypothesis is

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

Other elements of the score such as \( \frac{1}{T} \sum_{t=1}^{T} x_{it}^{(2)} (y_{it} - \mathbf{x}_{it}' \tilde{\beta}) \) and \( \frac{1}{T} \sum_{t=1}^{T} (y_{it} - \mathbf{x}_{it}' \tilde{\beta}) \) are equal to zero whether or not the null hypothesis is true and are, consequently, not included in \( \hat{g}_i(\tilde{\beta}) \).

Similar functions exist for nonlinear estimators:

\[ \text{IVQR: } \hat{g}_i(\beta) = \frac{1}{T} \sum_{t=1}^{T} z_{it}^{(k)} \left[ 1(y_{it} \leq \mathbf{x}_{it}' \beta) - \tau \right] \]

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

2.2 Inference with Asymptotically Uncorrelated Clusters

Initially, I discuss the inference procedure for uncorrelated clusters. Accounting for correlations across clusters will be discussed in Section 3 and involves a straightforward adjustment
which generates asymptotically independent\textsuperscript{2} clusters. Once this adjustment is made, the method discussed in this section can be used. The method uses elements of the score for each independent cluster:

\[
    g_i(\hat{\theta}) \equiv \frac{1}{T} \sum_{t=1}^{T} g_{it}(\hat{\theta})
\]

I make restrictions on \(g_i(\hat{\theta})\) which are assumed to hold under the null hypothesis. The extremum estimator setup (equation (2)) incorporates a wide range of estimation techniques. The main requirement necessary for implementation of CAI is that there exists a function or functions equal to zero under the null hypothesis and not equal to zero otherwise. For most estimation techniques, this is equivalent to an identification assumption.

\textbf{A1 (Identification):} There exists a \(1 \times H\) vector \(g_i(\cdot)\) such that

\[
    \begin{align*}
    E\left[g_i(\hat{\theta})\right] &= 0 & \text{if } a(b) = 0 \\
    E\left[g_i(\theta)\right] &= (q_1, \ldots, q_H)' & \text{where } q_h \neq 0 \text{ for all } h & \text{if } a(b) \neq 0
    \end{align*}
\]

This assumption rules out extremum estimators without a well-defined gradient such as the maximum score estimator (Manski [1975]). However, it includes GMM estimators and a wide range of other types of estimators when conditions for identification are met. As discussed previously, \textbf{A1} requires use of functions which are non-zero when the null hypothesis is not true.

The CAI procedure requires that a central limit theorem hold for each cluster \(i\) such that the score function converges to a normal distribution as \(T \to \infty\). I assume appropriate mixing conditions which imply a CLT (see Theorem 5.20 in White [2001]).

\textbf{A2 (Central Limit Theorem):} If \(a(b) = 0\), then

\begin{enumerate}
    \item \(\left\{g_{it}(\hat{\theta})\right\}\) is a strong mixing sequence in \(t\) with \(\alpha\) of size \(-\frac{r}{r-2}\), \(r > 2\).
    \item \(E\left|g_{ith}(\hat{\theta})\right|^{2r} < \Delta < \infty\) for all \(i, t, h\).
    \item \(\text{Var}\left(T^{-\frac{1}{2}} \sum_{t=1}^{T} g_{it}(\hat{\theta})\right)\) is uniformly positive definite with constant limit \(V_i\) for all \(i\).
\end{enumerate}

\textsuperscript{2}Because the inference procedure relies on functions which converge to the normal distribution, I use “independent” and “uncorrelated” interchangeably throughout the paper.
Assumption A2c allows for heterogeneity across clusters as there are no restrictions on variation in $V_i$ across clusters. The CAI method simulates the distribution of $g(\tilde{\theta})$ under the null hypothesis by perturbing the cluster-specific scores $g_i(\tilde{\theta})$. Let $\{W_i\}_{i=1}^N$ be an i.i.d. sample of weights, indexed by cluster. The weights perturb the cluster-specific scores and are independent of $g_i(\tilde{\theta})$. The weights satisfy the following conditions:

**A3 (Weights):** Assume i.i.d weights $\{W_i\}_{i=1}^N$ independent of $g_i(\tilde{\theta})$ such that for all $i$

a. $E[W_i] = 0$.


c. $E|W_i|^{2r} < \Delta < \infty$ for all $i$.

Given A2 and using weights that meet the conditions in A3, the first two moments of the perturbed score 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 $E[W_i^2] = 0$ and $E[W_i^4] = 1$ such that using this distribution is able to asymptotically match symmetric distributions (such as the normal distribution) and provide asymptotic refinement. Condition A3c is necessary to ensure that a CLT holds for the weighted scores and is met by the Rademacher distribution.

Under assumptions A1-A3, a CLT holds for each $g_i(\tilde{\theta})$ and each $W_i g_i(\tilde{\theta})$ under the null hypothesis. This is stated more formally below:

**Lemma 2.1.** Assume $a(b) = 0$ and that A1-A3 hold with $T \to \infty$. Then,

\[
\frac{1}{\sqrt{T}} \sum_{t=1}^{T} g_{it}(\tilde{\theta}) \xrightarrow{d} N(0, V_i) \tag{6}
\]

\[
\frac{1}{\sqrt{T}} \sum_{t=1}^{T} W_i g_{it}(\tilde{\theta}) \xrightarrow{d} N(0, V_i) \tag{7}
\]

Equation (6) holds by a CLT which permits dependence and non-identical random variables. See Theorem 5.20 in White [2001]. Because $W_i$ and $g_i(\tilde{\theta})$ are independent and $E[W_i] = E[g_i(\tilde{\theta})] = 0$,

$$\text{Var}(W_i g_i(\tilde{\theta})) = \text{Var}(W_i) \text{Var}(g_i(\tilde{\theta})) = \text{Var}(g_i(\tilde{\theta})).$$
The weights are constant within each cluster, preserving the within-cluster covariances and
permitting arbitrary within-cluster correlations. Given the result in equation (6), only a
restriction on the weights is necessary such that the conditions for a CLT still hold with
the perturbed scores, represented by equation (7). This condition was stated in A3c. See
Appendix A for a further discussion of Lemma 2.1.

Finally, I assume that the scores across clusters are asymptotically independent:

\( A4 \) (Asymptotic Independence): \( \text{Cov} \left( g_i(\tilde{\theta}), g_j(\tilde{\theta}) \right) \xrightarrow{p} 0 \) for all \( i \neq j \) as \( T \to \infty \).

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. The adjustment
discussed in Section 3 will create functions that meet \( A4 \) even in the presence of correlated
clusters. Given asymptotic independence and Lemma 2.1, it is straightforward to consider
the joint asymptotic distribution of the cluster-specific scores. The joint distribution of the
perturbed scores converges to the same distribution.

**Lemma 2.2.** Assume \( a(b) = 0 \) and that \( A1-A4 \) hold with \( T \to \infty \). Then,

\[
\sqrt{T} \begin{pmatrix}
  g_1(\tilde{\theta}) \\
  \vdots \\
  g_N(\tilde{\theta})
\end{pmatrix} \xrightarrow{d} N \left( \begin{pmatrix}
  V_1 & 0 \\
  0 & \ddots \\
  0 & 0 & V_N
\end{pmatrix} \right)
\]

\[
\sqrt{T} \begin{pmatrix}
  W_1 g_1(\tilde{\theta}) \\
  \vdots \\
  W_N g_N(\tilde{\theta})
\end{pmatrix} \xrightarrow{d} N \left( \begin{pmatrix}
  V_1 & 0 \\
  0 & \ddots \\
  0 & 0 & V_N
\end{pmatrix} \right)
\]

Lemma 2.2 follows from Lemma 2.1 and asymptotic independence. The variance
matrix in Lemma 2.2 is block diagonal. The importance of Lemma 2.2 is that both vectors
converge to the same distribution. Because \( E[W_iW_j] = 0 \), the convergence to the same
distribution can only occur if \( g_i(\tilde{\theta}) \) and \( g_j(\tilde{\theta}) \) are asymptotically uncorrelated since any
correlation would not be preserved by the weights. Given the result in Lemma 2.2, it is
possible to simulate the joint distribution of \( \left( g_1(\tilde{\theta}), \ldots, g_N(\tilde{\theta}) \right) \) by perturbing the cluster-
specific scores. For inference, the method uses a Wald Statistic defined by:

\[
S = \left( \frac{1}{N} \sum_{i=1}^{N} g_i(\tilde{\theta}) \right) ' \tilde{\Sigma}(\tilde{\theta})^{-1} \left( \frac{1}{N} \sum_{i=1}^{N} g_i(\tilde{\theta}) \right)
\]
where \( \hat{\Sigma}(\tilde{\theta}) = \frac{1}{N-1} \sum_{i=1}^{N} \left( g_i(\tilde{\theta}) - \left( \frac{1}{N} \sum_{i=1}^{N} g_i(\tilde{\theta}) \right) \right)' \left( g_i(\tilde{\theta}) - \left( \frac{1}{N} \sum_{i=1}^{N} g_i(\tilde{\theta}) \right) \right) \). This approach does not require estimation of \( V_i \) while permitting arbitrary within-cluster correlations. Instead, it uses the variance of \( g_i(\tilde{\theta}) \) across clusters. This test statistic has similarities to those proposed in Ibragimov and Müller [2010] and Canay et al. [2014]. For those test statistics, the papers propose using the mean and variance of the estimated parameters across clusters. Here, the model does not need to be 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.

To simulate the distribution of the test statistic, consider weights \( W_i^{(k)} \) which satisfy condition \( A3 \).

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

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

The scores are perturbed such that the test statistic can be simulated without re-estimating the model. Under the null hypothesis, \( g_i(\tilde{\theta}) \) should be centered around zero such that large values (in magnitude) suggest that the null hypothesis is incorrect. \( W_i g_i(\tilde{\theta}) \) is centered around zero whether the null hypothesis is true or not. Consequently, when the null hypothesis is not true, the large value of \( S \) should be rare in the distribution of \( S^{(k)} \).

**Theorem 2.1.** Assume \( a(b) = 0 \) and that \( A1-A4 \) hold with \( T \to \infty \). Then, \( S \) and \( S^{(k)} \) converge to the same distribution.

This result follows from Lemma 2.2 and the continuous mapping theorem. From Lemma 2.2, we know that the joint distribution of \( \left( g_1(\tilde{\theta}), \ldots, g_N(\tilde{\theta}) \right) \) is asymptotically equivalent to \( \left( W_1 g_1(\tilde{\theta}), \ldots, W_N g_N(\tilde{\theta}) \right) \). We can apply the continuous mapping theorem to prove that the corresponding test statistics also converge to the same distributions. Consequently, it is possible to simulate the distribution of the test statistic using weights satisfying \( A2 \). A further discussion is included in Appendix A.

Define the p-value as \( \hat{p} \) using the following formula and \( K \) simulations of the test
statistic:
\[ \hat{p} = \frac{1}{K} \sum_{k=1}^{K} 1(S > S^{(k)}) \]  

(10)

Corollary 2.2. Assume \( a(b) = 0 \) and that \( A1-A4 \) hold with \( T \to \infty \).

\[ P(\hat{p} \leq \alpha) \to \alpha \quad \text{for all} \quad \alpha \in [0, 1] \]  

(11)

Corollary 2.2 implies that we can generate appropriate p-values for any significance level through perturbing the test statistic.

2.3 Weights

As discussed earlier, the Rademacher distribution meets condition \( A2 \) with the additional advantage in finite samples that it matches symmetric distributions. In simulations, I use the Rademacher distribution to generate the weights. This distribution is frequently used for the wild bootstrap with a similar motivation and is recommended in Cameron et al. [2008]. However, Webb [2013] points out that a 2-point distribution is limited in the number of independent simulated values of the test statistic that it can produce. With only 6 clusters, a 2-point distribution generates only 32 (= 2^{G-1}) unique test statistics. The p-values are not point-identified and may include large intervals for small \( G \). Webb [2013] 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 [2013] proposes the following 6-point distribution, each with probability equal to \( \frac{1}{6} \):

\[ -\sqrt{1.5}, \quad -1, \quad -\sqrt{0.5}, \quad \sqrt{0.5}, \quad 1, \quad \sqrt{1.5} \]  

(12)

For similar reasons as discussed by Webb [2013], 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 [2013] distribution for \( G < 10 \) and the Rademacher distribution otherwise.
2.4 Discussion

This section has provided a valid inference procedure given asymptotically independent clusters. The motivation of this paper is to introduce a method for inference with dependent clusters. This requires a simple extension discussed in the next section. However, there are merits to using the above approach even with independent clusters. The method is simple and only requires estimation of the restricted estimator which enforces the null hypothesis. It does not require estimation of the same parameter for each cluster, allowing estimation of common parameters, including time fixed effects, across clusters. Furthermore, it does not require estimation of the Hessian, which can be complicated for many estimators. It also does not require additional assumptions concerning the homogeneity of the scores or the Hessian matrices across clusters, and there are no assumptions about the relative sizes of each cluster.

3 Asymptotically Correlated Clusters

Most applied work assumes independence across clusters. However, there are many reasons that this assumption may be violated. The inclusion of time fixed effects generates a mechanical correlation between clusters. Furthermore, some work has recognized that units that are geographically close may experience common shocks such that the units are not independent. This spatial clustering must be modeled and rules out dependence due to non-geographic factors such as similar industry compositions or political environments. With panel data, however, it is possible to observe and estimate dependence across clusters over time, isolate the independent components of each clusters, and then use inference procedures designed for asymptotically independent clusters.

3.1 Model

There are $H$ conditions as defined in $A1$. I index each by $h$ and consider each separately. Assume that under the null hypothesis, for each cluster $i$,

$$g_{ih} (\tilde{\theta}) = \sum_{\ell=1}^{L} a_{ih}^\ell M_{th}^\ell, \quad \text{with} \quad E[M_{th}^\ell] = 0 \quad \text{for all} \ \ell, \quad E[M_{th}^\ell M_{th}^k] = 0 \quad \text{for} \ \ell \neq k. \quad (13)$$

Given that no restrictions are placed on $a_{ih}^\ell$, this assumption is not restrictive in
terms of constraining the variance of $g_i(\tilde{\theta})$. Furthermore, no restrictions are placed on the relationship between $M_{th}^t$ and $M_{sh}^t$, permitting arbitrary within-cluster correlations. This equation also allows for correlations across clusters. The only binding restriction is that equation (13) rules out that lags or leads of the $M$ random variables independently affect some clusters and not others. This assumption is relaxed in Section 3.3.4, though it is not possible to allow for arbitrary correlations across clusters. However, this setup does allow for dependence across all observations given that $M_{th}^t$ and $M_{sh}^t$ are (possibly) correlated and potentially affect the score for each $i$ (i.e., $a_{th}^t \neq 0$). Since there are no restrictions on the $a_{th}^t$ terms, these correlations are “arbitrary” in the sense that there are no assumptions about the strength of dependence across clusters.

We can rewrite equation (13) in a triangular framework:

**A5:** $g_{ith}(\tilde{\theta})$ can be expressed as

$$g_{ith}(\tilde{\theta}) = b_{i+1,h}^ig_{i+1,th}(\tilde{\theta}) + \cdots + b_{Nh}^ig_{Nth}(\tilde{\theta}) + \mu_{ith}, \quad (14)$$

with $E[\mu_{ith}|g_{i+1,t}(\tilde{\theta}), \ldots, g_{Nt}(\tilde{\theta})] = 0$ and $E[\mu_{ith}\mu_{jsh}] = 0$ for all $s$ and $t$ and $i \neq j$.

For cluster $N$, this setup models $g_{Nth}(\tilde{\theta}) = \mu_{Nth}$. $\mu_{ith}$ represents the independent component of each cluster for time period $t$ and condition $h$. This specification is not restrictive in the sense that it is simply modeling the score for each cluster as an independent component ($\mu_{ith}$) and a dependent component ($b_{i+1,h}^ig_{i+1,th}(\tilde{\theta}) + \cdots + b_{Nh}^ig_{Nth}(\tilde{\theta})$). Modeling the relationship as linear is also not restrictive since we are only interested in the covariances between clusters. The implication of equation (13) is clear here - only period $t$ values of the score functions for other clusters enter the specification. Equation (14) assumes that lags and leads for other clusters are not independently related to $g_{ith}(\tilde{\theta})$. Note that this setup still permits correlations between $g_{it}(\tilde{\theta})$ and $g_{js}(\tilde{\theta})$ as long as that correlation acts through $g_{jt}(\tilde{\theta})$ or $g_{is}(\tilde{\theta})$. This assumption is met in the linear fixed effects case where $\epsilon_{it}$ and $\epsilon_{js}$ (see equation 1) are uncorrelated. By allowing for correlations, CAI also includes a wide range of cases where $\epsilon_{it}$ and $\epsilon_{js}$ are correlated. Furthermore, in Section 3.3.4, I discuss extensions of this model to include leads and lags.

The goal is to create score functions which are independent across clusters. It is straightforward to estimate equation (14) for each cluster $i$ and condition $h$ using OLS. Intuitively, the score functions of correlated clusters should co-vary. Estimation of (14) estimates these co-movements and allows one to subtract off the dependent component of
the score to isolate the independent component. With the estimates \( \hat{b}_{jh} \), it is possible to construct \( \hat{\mu}_{ith} \) and, subsequently,

\[
\hat{s}_{ih}(\hat{\theta}) = \frac{1}{T} \sum_{i=1}^{T} \hat{\mu}_{ith}.
\]

where \( \hat{s}_{ih}(\hat{\theta}) \) is the asymptotically-independent score function. In practice, equation (14) can be estimated with a constant. The constant is the estimate \( \hat{s}_{ih}(\hat{\theta}) \).

**Lemma 3.1.** Assume A1 and A5 hold and \( \hat{b}_{jh} \xrightarrow{p} b_{jh} \) for all \( i \) and \( j \). Then,

\[
\text{Cov} \left( g_{ith}(\tilde{\theta}) - \sum_{k=i+1}^{N} \hat{b}_{kh}^i g_{kth}(\tilde{\theta}) , g_{jsh}(\tilde{\theta}) - \sum_{k=j+1}^{N} \hat{b}_{kh}^j g_{ksh}(\tilde{\theta}) \right) \xrightarrow{p} 0 \ 	ext{for all} \ h, s, t.
\]

(15)

For \( T \to \infty \), OLS estimation provides consistent estimation of \( b_{jh}^i \) such that \( \hat{b}_{jh}^i \xrightarrow{p} b_{jh}^i \). The correlation adjustment simply involves regressing the score for a cluster on other scores. This is straightforward to implement. An immediate consequence of Lemma 3.1 is that A4 holds, implying that once the independent scores are generated, the procedure discussed in Section 2.2 can be used. Lemma 3.1 is discussed further in Appendix A.

### 3.2 Summary of CAI

The final inference procedure is summarized in the following steps.

1. Estimate \( \tilde{\theta} \) by imposing the null hypothesis (equation (5)).
2. Create \( g_i(\tilde{\theta}) \) for each \( i \).
3. Estimate equation (14) with OLS and create \( s_i(\tilde{\theta}) \).
4. Calculate \( S \) using \( s_i(\tilde{\theta}) \) and equation (8).
5. Simulate distribution of \( S \) using weights meeting condition A3 and equation (9) to generate \( S^{(1)}, \ldots, S^{(K)} \).
6. Generate p-value using (10).

In all simulations, I use \( K = 999 \).
3.3 Extensions and Modifications

3.3.1 Alternate Simulation of Distribution of $S$

To simulate the distribution of $S$, I proposed using equation (9). However, it is asymptotically equivalent to use the unconstrained estimate $\hat{\theta}$ to generate the simulated distribution:

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

(16)

where $\widehat{\Sigma}^{(k)}(\hat{\theta}) = \frac{1}{N-1} \sum_{i=1}^{N} \left( W_i^{(k)} g_i(\hat{\theta}) - \left( \frac{1}{N} \sum_{i=1}^{N} W_i^{(k)} g_i(\hat{\theta}) \right) \right)' \left( W_i^{(k)} g_i(\hat{\theta}) - \left( \frac{1}{N} \sum_{i=1}^{N} W_i^{(k)} g_i(\hat{\theta}) \right) \right)$. In simulations, I provide results using both methods to simulate the distribution of the test statistic.

3.3.2 Alternative Dependence Estimation

The individual estimates of $b_{i+1}^{i+1,h}$ are only used to create the independent score functions and are of interest in themselves. Alternatively, it is possible to model the dependence across scores using

$$
g_{ith}(\hat{\theta}) = c_i^h \left[ \frac{1}{N - i + 1} \sum_{j=i+1}^{N} g_{jth}(\hat{\theta}) \right] + \mu_{ith},
$$

This approach only requires estimation of one parameter per unit where $c_i^h$ represents the aggregate correlation between the unit $i$ score and the scores for units $i < j \leq N$. While CAI depends on large $T$, estimating the above equation may have better properties in finite sample when $T$ is not much larger than $N$.

3.3.3 Large $N$, Small $T$

This paper discusses inference for large $T$. However, since CAI adjusts for correlations across clusters, the procedure is also appropriate for small $T$ and large $N$. In that case, valid inference requires assuming that each time period is a cluster and then adjusting for correlations across time periods (i.e., within geographic units). Equation (14) adjusts for within-unit correlations in the small $T$ case. Conceptually, since the CAI approach allows
for arbitrary within-unit dependence and arbitrary correlations across time, there is little difference whether “clusters” are defined by time or unit as long as a CLT holds in the one dimension.

3.3.4 Lags and Leads

The only restriction implicit in equation (14) is that correlations across clusters act through the scores in the same time period. Because arbitrary within-cluster correlations are allowed, this assumption permits dependence across all observations. A4 rules out the case, as an example, where the score for unit $i$ in time period $t - 1$ has an independent effect on unit $j$ in time period $t$ that does not act through the score of unit $i$ in period $t$. This assumption can be relaxed by including lagged values (or leads) in equation (14). With lags, I model the scores using $AR(p)$ notation:

$$g_{ith}(\tilde{\theta}) = \sum_{s=0}^{p} b_{i,s}^{j} g_{i+1,t-s,h}(\tilde{\theta}) + \cdots + \sum_{s=0}^{p} b_{N,s}^{i} g_{N,t-s,h}(\tilde{\theta}) + \mu_{ith}.$$ 

Leads can also be included. The variable $p$ controls the length of the lags and, while constant in the equation above, can vary by $i$. It is not possible to allow for arbitrary correlations across clusters such that the scores for each time period independently affect the scores for each time period of another cluster. However, equation (14) can be relaxed substantially to allow for more flexibility.

3.4 Cluster Covariance Estimator

Section 2.2 introduced an inference procedure that is valid for asymptotically independent clusters. Section 3 discusses adjustments for correlated clusters and merges this adjustment with the Section 2.2 inference procedure. However, the broader point made in this section is that panel data allow for estimation of correlations across clusters. This insight can be used with other inference methods such as CCE.

Under condition A5 and assumptions discussed in BCH or Hansen [2007], CCE can produce valid inference with correlated clusters. I focus on the method discussed in Section 2.2 primarily because it relaxes many of the assumptions that CCE would require when the number of clusters is small. However, given CCE’s popularity, it is important to recognize that the correlation-adjustment proposed throughout Section 3 is also valid for CCE.
4 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 show that CAI works well relative to the methods used in the literature. I will also provide simulations with highly-dependent clusters in which conventional methods perform especially poorly. In each set of simulations, I include rejections rates using CCE. I provide three CCE reject rates - (1) assuming each unit is a cluster ($G = N$); (2) using $G = 4$; (3) using $G = 2$. The latter two cases are suggested by BCH. In all cases, I use a $t$-distribution with $G - 1$ degrees of freedom (as proposed in BCH). For some simulations, it is clear how to assign units into each clusters. In other cases, however, I make this assignment randomly (without loss of generality since there is no spatial dependence in those simulations). Finally, I also include rejection rates using the wild cluster bootstrap which simulates the distribution of a Wald statistic. This approach also imposes the null hypothesis. Cameron et al. [2008] finds that this approach has appropriate rejection rates even for small $G$. When $G < 10$, I will use the Webb [2013] weights, though Cameron et al. [2008] uses the Rademacher distribution.

I provide two different rejection rates for CAI as discussed in Section 3.3.1. First, I include rejection rates in which the test statistic is simulated using the restricted estimate. Second, I include rejection rates in which the test statistic is simulated using the unrestricted estimate. I expect both should reject at the appropriate rate.

4.1 BCH Replication

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

$$\ln(y_{st}) = \beta x_{st} + \alpha_s + \alpha_t + \epsilon_{st},$$  \hspace{1cm} (17)

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.
in the following manner:

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

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 \( \gamma \) and \( \rho_j \), respectively. I use the same data and simulate policies in the same manner.

Following BCH, I use \( \gamma = 0.8 \) and set \( \sigma \) such that the standard deviation of \( x_{st} \) and \( \ln(y_{st}) \) are the same. BCH consider two data generating processes. First, they consider an AR(13) process with \( \rho_1 = 0.95, \rho_{12} = 0.68, \) and \( \rho_{13} = -0.68 \). Second, they consider 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 by Census regions),\(^3\) and 2 clusters (dividing the country into two).\(^4\) In each case, I use a \( t \)-distribution with \( G - 1 \) degrees of freedom where \( G \) is the number of clusters. I also report rejection rates using a wild bootstrap. Finally, I report two sets of rejection rates using the inference procedure developed in this paper. I simulate the test statistic distribution using either the restricted estimates ("Null") or the estimates from the unrestricted model ("Estimated") (as discussed in 3.3.1).

BCH primarily uses first-differences to estimate (17) 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 due to spatial dependence in \( x \) (by design) and the outcome variable (due to correlated, unobserved shocks). The wild bootstrap also overrejects. Dividing the

\(^3\)I follow BCH in reassigning some states to make the clusters approximately equal-sized. See BCH footnote 23.

\(^4\)Again, I follow BCH to assign states to East or West. See BCH footnote 23.
Table 1: Simulation Results from BLS Data

<table>
<thead>
<tr>
<th></th>
<th>First-Differences</th>
<th>Fixed Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AR(13)</td>
<td>AR(1)</td>
</tr>
<tr>
<td>CCE, G=49</td>
<td>0.150</td>
<td>0.145</td>
</tr>
<tr>
<td>CCE, G=4</td>
<td>0.069</td>
<td>0.058</td>
</tr>
<tr>
<td>CCE, G=2</td>
<td>0.059</td>
<td>0.055</td>
</tr>
<tr>
<td>Wild Bootstrap</td>
<td>0.163</td>
<td>0.151</td>
</tr>
<tr>
<td>CAI, Null</td>
<td>0.050</td>
<td>0.052</td>
</tr>
<tr>
<td>CAI, Estimated</td>
<td>0.051</td>
<td>0.055</td>
</tr>
</tbody>
</table>

Notes: Table reports rejection rates for 5% level tests from simulations using BLS unemployment data. All results are based on 1000 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. CAI refers to the correlation-adjusted score inference procedure introduced in this paper. Sample size is $N=49$, $T=383$.

states into 4 clusters improves the rejection rate. Dividing all states into 2 clusters further improves the rejection rate to 0.059. The CAI procedure fares even better with rejection rates of 0.050 and 0.051, depending on which scores are used to simulate the distribution of possible scores.

The next column presents the equivalent results for the AR(1) process. The results are generally similar. Next, I estimate equation (17) using state and time fixed effects. CCE fares worse when fixed effects are used instead of first-differences and overrejects at a higher rate. The CAI procedure also rejects more often, though it is still close to the expected rate. Similar patterns for the other inferences methods are found for the AR(1) process, but the CAI procedure underrejects slightly in this case.

4.2 CGM Replication

Next, I replicate simulations provided in Cameron et al. [2008]. The data are generated in the following manner:

\[
y_{st} = \beta x_{st} + u_{st},
\]
\[
x_{st} = z_s + z_{st},
\]
\[
u_{st} = \epsilon_s + \epsilon_{st},
\]
where $z_s$, $z_{st}$, $\epsilon_s$, and $\epsilon_{st}$ are independent and distributed $N(0,1)$. I vary the number of states (indexed by $s$). When the number of states is less than 10, I use the 6-point wild bootstrap proposed in Webb [2013], instead of the 2-point bootstrap used in Cameron et al. [2008]. I include simulations where the number of states is equal to 6, 10, 30, and 50. I use even numbers so that the BCH method for $G = 2$ can have an equal number of states in each cluster (an assumption made in BCH). Since there is no cross-sectional dependence, I randomize which states are included in each cluster.

Table 2: Simulation Results Replicating CGM

<table>
<thead>
<tr>
<th>Number of States (N)</th>
<th>6</th>
<th>10</th>
<th>30</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCE, $G=N$</td>
<td>0.086</td>
<td>0.072</td>
<td>0.059</td>
<td>0.060</td>
</tr>
<tr>
<td>CCE, $G=4$</td>
<td>0.080</td>
<td>0.051</td>
<td>0.057</td>
<td>0.054</td>
</tr>
<tr>
<td>CCE, $G=2$</td>
<td>0.058</td>
<td>0.053</td>
<td>0.046</td>
<td>0.045</td>
</tr>
<tr>
<td>Wild Bootstrap</td>
<td>0.099</td>
<td>0.069</td>
<td>0.051</td>
<td>0.055</td>
</tr>
<tr>
<td>CAI, Null</td>
<td>0.059</td>
<td>0.055</td>
<td>0.039</td>
<td>0.054</td>
</tr>
<tr>
<td>CAI, Estimated</td>
<td>0.046</td>
<td>0.044</td>
<td>0.038</td>
<td>0.053</td>
</tr>
</tbody>
</table>

Notes: Table reports rejection rates for 5% level tests from simulations. All results are based on 1000 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. CAI refers to the correlation-adjusted score inference procedure introduced in this paper.

Table 2 presents the simulation results. The BCH approach fares well even for $N = 10$ when $G = 2$ or $G = 4$. The wild bootstrap overrejects when $N \leq 10$. CAI performs relatively well for all $N$.

I also use the data generated above and estimate the specification using median regression to illustrate the usefulness of the proposed inference procedure for nonlinear models. 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 used in Cameron et al. [2008] is only useful for linear models so I do not include any wild bootstrap results for the median regression simulations.

Table 3 presents the simulation results. CCE with only 2 clusters performs well for all $N$. CAI also rejects at close to the appropriate rates.

Next, I generate similar data but also include state and time fixed effects. I estimate the specification using OLS with state and time fixed effects given that $x$ is a function of
Table 3: Median Regression Simulations

<table>
<thead>
<tr>
<th>Number of States (N)</th>
<th>6</th>
<th>10</th>
<th>30</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCE, G=N</td>
<td>0.067</td>
<td>0.073</td>
<td>0.060</td>
<td>0.065</td>
</tr>
<tr>
<td>CCE, G=4</td>
<td>0.072</td>
<td>0.061</td>
<td>0.054</td>
<td>0.051</td>
</tr>
<tr>
<td>CCE, G=2</td>
<td>0.041</td>
<td>0.052</td>
<td>0.041</td>
<td>0.050</td>
</tr>
<tr>
<td>CAI, Null</td>
<td>0.063</td>
<td>0.051</td>
<td>0.053</td>
<td>0.049</td>
</tr>
<tr>
<td>CAI, Estimated</td>
<td>0.055</td>
<td>0.046</td>
<td>0.054</td>
<td>0.051</td>
</tr>
</tbody>
</table>

Notes: Table reports rejection rates for 5% level tests from simulations. All results are based on 1000 simulations. CCE refers to the cluster covariance matrix estimator. A $t$-distribution with $G - 1$ degrees of freedom is used. CAI refers to the correlation-adjusted score inference procedure introduced in this paper.

both sets of fixed effects. The motivation for these simulations is the difficulty that fixed effects can cause for many inference procedures. The data are generated in the following manner:

$$y_{st} = \beta x_{st} + 5\alpha_s + 5\gamma_t + u_{st},$$

$$x_{st} = 5\alpha_s + 5\gamma_t + z_s + z_{st},$$

$$u_{st} = \epsilon_s + \epsilon_{st},$$

where $\alpha_s$, $\gamma_t$, $z_s$, $z_{st}$, $\epsilon_s$, and $\epsilon_{st}$ are independent and distributed $N(0, 1)$. The only changes from the previous simulated data are the inclusion of state and time fixed effects. The results are presented in Table 4.

Table 4: Simulation Results Replicating CGM with Fixed Effects

<table>
<thead>
<tr>
<th>Number of States</th>
<th>6</th>
<th>10</th>
<th>30</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCE, G=N</td>
<td>0.069</td>
<td>0.069</td>
<td>0.058</td>
<td>0.049</td>
</tr>
<tr>
<td>CCE, G=4</td>
<td>0.074</td>
<td>0.064</td>
<td>0.071</td>
<td>0.054</td>
</tr>
<tr>
<td>CCE, G=2</td>
<td>0.063</td>
<td>0.054</td>
<td>0.052</td>
<td>0.047</td>
</tr>
<tr>
<td>Wild Bootstrap</td>
<td>0.072</td>
<td>0.072</td>
<td>0.053</td>
<td>0.051</td>
</tr>
<tr>
<td>CAI, Null</td>
<td>0.052</td>
<td>0.045</td>
<td>0.050</td>
<td>0.057</td>
</tr>
<tr>
<td>CAI, Estimated</td>
<td>0.047</td>
<td>0.049</td>
<td>0.047</td>
<td>0.061</td>
</tr>
</tbody>
</table>

Notes: Table reports rejection rates for 5% level tests from simulations. All results are based on 1000 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. CAI refers to the correlation-adjusted score inference procedure introduced in this paper.
The fixed effects lead to overrejection for all methods using CCE when \( N \) is small. The wild bootstrap approach also overrejects. Overrejection for these inference methods was expected given that the fixed effects generate correlations across clusters. The CAI method fares well and rejects at the appropriate rate. By adjusting for the relationship between units, the method creates independent clusters and rejects at close to 5%.

4.3 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 data for unit \( i \) in group \( g \) at time \( t \). There are two units per group and I generate data such that the units are highly-correlated.

\[
y_{igt} = \beta x_{igt} + \gamma_t + \alpha_i + u_{igt},
\]
\[
x_{igt} = \gamma_t + \alpha_i + z_{gt} + 0.1 \times \nu_{igt},
\]
\[
z_{gt} = 0.9 \times z_{g,t-1} + \phi_{gt},
\]
\[
u_{igt} = u_{gt} + 0.1 \times \zeta_{igt},
\]
\[
u_{gt} = 0.9 \times u_{g,t-1} + \eta_{gt},
\]

where \( \phi_{gt}, \nu_{igt}, \zeta_{igt}, \eta_{gt}, \gamma_t, \) and \( \alpha_i \) 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 \( \omega \) are, respectively, common components of these random variables within each group. I do not assume that states within the same group are spatially close. Instead, 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 the BCH approach. 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.

Given that correlated states are not correctly placed in the same cluster for CCE, it is not surprising that the BCH method consistently and severely overrejects, as shown in Table 5. The wild bootstrap also does not account for the correlations across clusters and performs poorly. CAI, however, performs well even with \( N = 6 \). These simulations illustrate the power of the inference method introduced in this paper. The procedure estimates the relationship between units and uses only the independent variation for inference.
Table 5: Correlated States Simulation

<table>
<thead>
<tr>
<th>Number of States (N)</th>
<th>6</th>
<th>10</th>
<th>30</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCE, G=N</td>
<td>0.380</td>
<td>0.289</td>
<td>0.198</td>
<td>0.189</td>
</tr>
<tr>
<td>CCE, G=4</td>
<td>0.408</td>
<td>0.196</td>
<td>0.069</td>
<td>0.086</td>
</tr>
<tr>
<td>CCE, G=2</td>
<td>0.358</td>
<td>0.248</td>
<td>0.121</td>
<td>0.120</td>
</tr>
<tr>
<td>Wild Bootstrap</td>
<td>0.367</td>
<td>0.278</td>
<td>0.198</td>
<td>0.184</td>
</tr>
<tr>
<td>CAI, Null</td>
<td>0.058</td>
<td>0.056</td>
<td>0.043</td>
<td>0.045</td>
</tr>
<tr>
<td>CAI, Estimated</td>
<td>0.052</td>
<td>0.054</td>
<td>0.043</td>
<td>0.045</td>
</tr>
</tbody>
</table>

Notes: Table reports rejection rates for 5% level tests from simulations. All results are based on 1000 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. CAI refers to the correlation-adjusted score inference procedure introduced in this paper.

4.4 Power Curves

Using the data-generating process described in Section 4.1, I study rejection rates when $\beta \neq 0$. I use the AR(13) process and include state fixed effects. I set $\beta$ equal to a constant and generate a new outcome variable, adding $\beta x_{st}$ to $\ln(y_{st})$. Figure 1 shows the rejection rate for each constant. I compare the power curve of the correlation adjusted score procedure to the rejection rate using CCE with $G = 2$.

The rejection rate for CAI is always higher than the CCE rejection rate except for the $\beta = 0$ case. This may not be surprising as BCH allows for arbitrary within-cluster dependence where each cluster includes half of the units. CAI models the dependence across units more explicitly.

5 Discussion and Conclusion

This paper introduces a simple method to conduct inference with correlated, heterogeneous clusters and is valid even when the number of cluster 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 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 clusters and heterogeneity in both the score and
Figure 1: Power Curve, BCH Data, AR(13) Process with State Fixed Effects

Hessian functions.

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 estimated and inference can proceed as if the clusters were uncorrelated. This method should be useful generally for linear and nonlinear estimators with panel data.
A Appendix

Lemma 2.1. Assume $a(b) = 0$ and that $A1$-$A3$ hold with $T \to \infty$. Then,

\[
\frac{1}{\sqrt{T}} \sum_{t=1}^{T} g_{it}(\tilde{\theta}) \xrightarrow{d} N(0, V_i) \quad (6)
\]

\[
\frac{1}{\sqrt{T}} \sum_{t=1}^{T} W_i g_{it}(\tilde{\theta}) \xrightarrow{d} N(0, V_i) \quad (7)
\]

Proof. Equation (6) follows from a standard CLT (see Theorem 5.20 in White [2001]).

Given equation (6), it is straightforward to show equation (7) holds. First, given the conditions met by the weights, $g_i(\tilde{\theta})$ and $W_i g_i(\tilde{\theta})$ have the same asymptotic variance

\[
\text{Var}(W_i g_i(\tilde{\theta})) = \text{Var}(W_i) \text{Var}(g_i(\tilde{\theta})) = \text{Var}(g_i(\tilde{\theta})).
\]

Second, by $A2b$ and $A3c$,

\[
E \left| W_i g_i(\tilde{\theta}) \right|^r \leq E \left| W_i \right|^r \left| g_i(\tilde{\theta}) \right|^r < \Delta < \infty.
\]

Lemma 2.2. Assume $a(b) = 0$ and that $A1$-$A4$ hold with $T \to \infty$. Then,

\[
\sqrt{T} \begin{pmatrix} g_1(\tilde{\theta}) \\ \vdots \\ g_N(\tilde{\theta}) \end{pmatrix} \xrightarrow{d} N \begin{pmatrix} V_1 & 0 \\ 0 & V_N \end{pmatrix}
\]

\[
\sqrt{T} \begin{pmatrix} W_1 g_1(\tilde{\theta}) \\ \vdots \\ W_N g_N(\tilde{\theta}) \end{pmatrix} \xrightarrow{d} N \begin{pmatrix} V_1 & 0 \\ 0 & V_N \end{pmatrix}
\]

Proof. By $A4$, $\text{Cov} \left( g_i(\tilde{\theta}), g_j(\tilde{\theta}) \right) \xrightarrow{p} 0$. 

\[\text{□}\]
Given this result and $A_2$, $\text{Cov} \left( W_i g_i(\tilde{\theta}), W_j g_j(\tilde{\theta}) \right) \xrightarrow{p} 0$.

\[ \text{Theorem 2.1.} \text{ Assume } a(b) = 0 \text{ and that } A_1 - A_4 \text{ hold with } T \to \infty. \text{ Then, } S \text{ and } S^{(k)} \text{ converge to the same distribution.} \]

This result follows immediately from Lemma 2.2 and the continuous mapping theorem.

\[ \text{Lemma 3.1.} \text{ Assume } A_1 \text{ and } A_5 \text{ hold and } \hat{b}_{ih} \xrightarrow{p} b_{ih} \text{ for all } i \text{ and } j. \text{ Then,} \]

\[ \text{Cov} \left( g_{ih}(\tilde{\theta}) - \sum_{k=i+1}^{N} \hat{b}_{ik} g_{kh}(\tilde{\theta}) , g_{jsh}(\tilde{\theta}) - \sum_{k=j+1}^{N} \hat{b}_{ik} g_{kh}(\tilde{\theta}) \right) \xrightarrow{p} 0 \text{ for all } h. \]

\[ (15) \]

\[ \text{Proof. By } A_5, \]

\[ \text{Cov} \left( g_{ih}(\tilde{\theta}) - \sum_{k=i+1}^{N} \hat{b}_{ik} g_{kh}(\tilde{\theta}) , g_{jsh}(\tilde{\theta}) - \sum_{k=j+1}^{N} \hat{b}_{ik} g_{kh}(\tilde{\theta}) \right) \xrightarrow{p} \text{Cov} \left( \mu_{ih}, \mu_{jsh} \right) = 0 \]
References


Timothy G Conley and Christopher R Taber. Inference with “difference in differences” with


