Inference with Dependent Data Using Cluster Covariance Estimators*

C. Alan Bester,† Timothy G. Conley,‡ and Christian B. Hansen§

First Draft: February 2008. This Draft: November 2010.

Abstract. This paper presents an inference approach for dependent data in time series, spatial, and panel data applications. The method involves constructing t and Wald statistics using a cluster covariance matrix estimator (CCE). We use an approximation that takes the number of clusters/groups as fixed and the number of observations per group to be large. The resulting limiting distributions of the t and Wald statistics are standard t and F distributions where the number of groups plays the role of sample size. Using a small number of groups is analogous to 'fixed-b' asymptotics of Kiefer and Vogelsang (2002, 2005) (KV) for heteroskedasticity and autocorrelation consistent inference. We provide simulation evidence that demonstrates the procedure substantially outperforms conventional inference procedures.

Keywords: HAC, panel, robust, spatial

JEL Codes: C12, C21, C22, C23

*We would like to thank seminar participants at the Atlanta, Chicago, and Richmond Federal Reserves, Arizona State University, Boston College, Cincinatti, Concordia, Notre Dame, Purdue, Stanford University, Syracuse University, UCLA, University of Chicago Booth School, University of Iowa, University of Kentucky, University of Pennsylvania, University of Texas - Austin, and University of Western Ontario as well as conference participants at Yale, University of Montreal, and the Spatial Econometrics Association Meetings for helpful comments and discussion. We thank anonymous referees and the editors for valuable suggestions that greatly improved the paper. We are also grateful for support from the Neubauer Family Faculty Fund and from the IBM Corporation Faculty Research Fund at the University of Chicago Booth School of Business.

†University of Chicago Booth School of Business, cbester@chicagobooth.edu

‡University of Western Ontario, tconley3@uwo.ca

§University of Chicago Booth School of Business, chansen1@chicagobooth.edu

1. Introduction

Many economic applications involve dependent data. The study of serial dependence is fundamental in the analysis of time series, and cross-sectional or spatial dependence is an important feature in many types of cross-sectional and panel data. While the dependence structure in a given data set is often not the object of interest, it is well understood that inference about parameters of interest, such as regression coefficients, may be severely distorted when one does not account for this dependence. This paper presents a simple method for conducting inference about estimated parameters with spatially dependent data. This setup includes time series and panel data as special cases.

There are two main methods for conducting nonparametric inference with dependent data. By far the most common is to use a limiting normal approximation that depends on an unknown variance-covariance matrix. One then 'plugs-in' a covariance matrix estimator that is consistent under heteroskedasticity and autocorrelation of unknown form (commonly called a HAC estimator) in place of this unknown matrix. For time series econometrics, this plug-in HAC covariance matrix approach has been popular since at least Newey and West (1987) and for spatial econometrics it dates to Conley (1996, 1999).

Kiefer and Vogelsang (2002, 2005) (KV) propose an alternative to the conventional plug-in approach. KV consider the limiting properties of conventional time-series HAC estimators under an asymptotic sequence in which the HAC smoothing or cutoff parameter is proportional to the sample size, as opposed to the conventional sequence where the smoothing parameter grows more slowly than the sample size. Under the KV sequence, the HAC estimator converges in distribution to a non-degenerate random variable. KV provide approximate distributions for commonly-used test statistics accounting for this randomness in the HAC covariance estimator. Taking a t-statistic as an example, the conventional approach described in the previous paragraph views the denominator as consistent and its variability is not accounted for. In contrast, the KV approach treats the

t-statistic denominator as a random variable and thus uses a ratio of limiting random variables as a reference distribution. The resulting limit distribution for the t-statistic is pivotal but nonstandard, so critical values are obtained by simulation. KV provide convincing simulation evidence that their approximation outperforms the plug-in approach. Jansson (2004) and Sun, Phillips, and Jin (2008) show formally that the 'fixed-b' approximation is a refinement of the standard asymptotic approximation in Gaussian location models.

In this paper, we present a simple method for conducting inference in the spirit of KV that also applies to spatially dependent and panel data. As in KV, we calculate limiting distributions for common test statistics viewing covariance estimators as random variables in the limit. We differ from KV in the type of covariance estimator we employ. Our methods use what is commonly called a cluster covariance matrix estimator (CCE) which is popular in applied microeconomics. Under conditions on group structure and dependence across observations, we derive the behavior of test statistics formed using the CCE. We obtain results under asymptotics that treat the number of groups as fixed and the number of observations within a group as large. Under this approximating sequence, t- and Wald statistics follow standard t- and F distributions with degrees of freedom determined by the number of groups used in constructing the CCE.[1]

Cluster covariance estimators are routinely used with data that has a group structure with independence assumed across groups.[2] Typically, inference is conducted in such settings under the assumption that there are a large number of these independent groups. In economic applications, data often feature natural groupings, such as firm outcomes in a given year or household outcomes in a given census tract. In many cases, observations in different groups are *not* independent; for

[1]In a working version of this paper, Bester, Conley, and Hansen (2010), we also present consistency results for the CCE without assuming independence between observations within different groups when both the number of groups and their size are allowed to grow at certain rates.

[2]See Wooldridge (2003) for a concise review of this literature. See also Liang and Zeger (1986), Arellano (1987), Bertrand, Duflo, and Mullainathan (2004), and Hansen (2007).

example, consider firms in the same industry in subsequent years, or households in two adjacent census tracts. However, with enough weakly dependent data, we show that groups can be chosen by the researcher so that group-level averages are approximately independent. Intuitively, if groups are large enough and well-shaped (e.g. do not have gaps), the majority of points in a group will be far from other groups, and hence approximately independent of observations from other groups provided the data are weakly dependent. The key prerequisite for our methods is the researcher's ability to construct groups whose averages are approximately independent. As we show later, this often requires that the number of groups be kept relatively small, which is why our main results explicitly consider a fixed (small) number of groups.

We note that the idea of partitioning the data into researcher-defined groups to overcome dependence problems has a long history in econometrics and statistics. In time series analysis, the idea dates to at least Bartlett (1950) who discusses partitioning a time series into a set of 'short' series and averaging across these short series to approximate averages across independent series. Specifically, he mentions that one might use a cluster estimator formed by calculating periodograms for each short series and then averaging across these 'short periodograms' to obtain a spectral density estimator. This estimator is essentially the same as the CCE. Bartlett notes that the cluster estimator obviously does not use information on covariances from pairs of observations located in different short series and motivates what has become known as a Bartlett spectral density estimator as a modification of the cluster estimator to include these omitted terms. The idea of blocking or partitioning data is also important in the literature on bootstrapping under dependence; see, for example, Lahiri (2003). The widely-used Fama and MacBeth (1973) procedure consists of basing inference on a set of (approximately) independent point estimates, each from one element of a partition of a dataset. A recent paper by Ibragimov and Müller (2006) (IM) provides a formal treatment of the Fama-Macbeth procedure, focusing upon properties of t-tests using these sets of point estimates. IM note that the key high-level condition required for such tests' validity is having a set of groups whose averages are asymptotically independent. In our paper, we provide a set

of primitive conditions for this high-level assumption to be satisfied in a spatial (vector-indexed) context which can immediately be used to establish the validity of the IM procedure for conducting inference with spatial dependence.[3]

Our results concern the behavior of the usual t-statistics and Wald tests formed using the CCE as a covariance matrix under limits corresponding to a fixed number of groups, each of which consists of a large number of observations. We show that Wald statistics follow F-distributions and t-statistics follow t-distributions in the limit up to simple and known scale factors that depend only on the number of groups used in forming the CCE and the number of restrictions being tested. Our regularity conditions involve moment and mixing rate restrictions, weak homogeneity assumptions on second moments of regressors and unobservables across groups, and restrictions on group boundaries. These moment and mixing conditions are implied by routine assumptions necessary for use of central limit approximations and the required homogeneity is less restrictive than covariance stationarity.

Our theoretical results also contribute to the growing literature on inference with spatial data; that is, data in which dependence is indexed in more than one dimension. Examples of papers in this literature are Conley (1996, 1999), Kelejian and Prucha (1999, 2001), Lee (2004, 2007a, 2007b), and Jenish and Prucha (2007). This paper also complements Bester, Conley, Hansen, and Vogelsang (2008) (BCHV) which extends the KV approach to conventional spatial HAC estimators. As in KV, the reference distributions obtained in BCHV are pivotal but nonstandard and critical values must be obtained by simulation. Also, the reference distributions in BCHV explicitly depend on the shape of the sample space. Relative to Bester, Conley, Hansen, and Vogelsang (2008), we consider the CCE rather than a HAC estimator. The CCE has a number of appealing features relative to a more conventional HAC estimator. It is simple to implement and is very widely used in empirical economic research. Providing formal conditions under which inference based on the CCE remains

[3]We provide a more complete discussion of our procedure relative to IM near the end of Section 3.1 and simulation evidence regarding their relative performance in Section 4.3.

valid and a procedure which has good size properties in very general settings is a main contribution of this paper. The homogeneity conditions under which the CCE provides valid inference are also somewhat weaker than those used to establish the results in KV and Bester, Conley, Hansen, and Vogelsang (2008). In our simulations, HAC-based tests with KV critical values do have somewhat larger size distortions than the CCE-based tests. The drawback of using the CCE appears to be a small loss in power relative to tests based on conventional HAC estimators using KV critical values.

We present simulation evidence on the performance of our estimator in time series, spatial, and panel data contexts. The time series setting and cross-sectional setting with spatial dependence use simulated treatments and outcomes. We also consider a panel context using actual unemployment rate outcomes regressed on simulated treatments. In time series and cross sectional settings, the simulation evidence clearly demonstrates that plug-in HAC inference procedures, which rely on asymptotic normal and χ^2 approximations, may suffer from substantial size distortions. In all cases, the simulations clearly illustrate that inference procedures that ignore either cross-sectional or temporal dependence, such as clustering based on only state or month in our unemployment simulations, are severely size distorted. We also provide simulation results comparing our approach to IM that demonstrate that neither procedure dominates the other. Overall, the simulations show that, provided the number of groups is small and correspondingly the number of observations per group is large, our proposed test procedure has actual size close to nominal size and non-negligible power.

The remainder of the paper is organized as follows. Section 2 presents estimators and notation for the linear regression model. Section 3 discusses the large sample properties of t and Wald statistics formed using the CCE. Section 4 presents simulation evidence regarding the tests' performance. Section 5 concludes. Proofs are relegated to the Appendix.

2. Methodology

For ease of exposition, we first present our method in the context of ordinary least squares (OLS) esimation of the linear model. An outline of the extension of our results to nonlinear models is given in the appendix.

2.1. Model and Notation

We use two sets of notation, corresponding to the model at the individual and group level. For simplicity we take individual observation i to be indexed by a point s_i. The regression model is

$$y_{s_i} = x'_{s_i}\beta + \varepsilon_{s_i}.$$

The variables y_{s_i} and ε_{s_i} are a scalar outcome and regression error, and x_{s_i} is a $k \times 1$ vector of regressors that are assumed orthogonal to ε_{s_i}. We use N to refer to the total number of observations.

We characterize the nature of dependence between observations via their indexed locations $s_1, ..., s_N$. This is routine for time series data where these indices reflect the timing of the observations. Following Conley's (1996, 1999) treatment of spatial dependence, we explicitly consider vector indices that allow for the complicated dependence structures found in spatially dependent data or space-time dependence in panel data. Locations provide a structure for describing dependence patterns.[4] The key assumption we make regarding dependence between observations is that they are weakly dependent, meaning that random variables approach independence as the distance between their locations grows. Observations at close locations are allowed to be highly correlated and correlation patterns within sets of observations can be quite complicated.

[4]The economics of the application often provides considerable guidance regarding the index space and metric. For example, when local spillovers or competition are the central economic features, obvious candidate metrics are measures of transaction/travel costs limiting the range of the spillovers or competition. Index spaces are not limited to the physical space or times inhabited by the agents and can be as abstract as required by the economics of the application; e.g., see Conley and Ligon (2002) and Pulvino (1998).

Our methods involve partitioning the data into groups defined by the researcher. We define G_N to be the total number of groups and index them by $g = 1, ..., G_N$. For simplicity, our presentation ignores integer problems and takes the groups to be of common size L_N. It will often be convenient to use group-level notation for the regression model. Let y_g be an $L_N \times 1$ vector defined by stacking each of the individual y_s within a group g, and likewise let ε_g be a stacked set of error terms and x_g be an $L_N \times k$ matrix with generic row x_s'. This yields a group level regression equation:

$$y_g = x_g\beta + \varepsilon_g.$$

The econometric goal is to conduct inference about β. We will examine the OLS estimator of β using the whole sample, which of course can be written as

$$\hat{\beta}_N = \left(\sum_{i=1}^{N} x_{s_i} x_{s_i}^{'}\right)^{-1} \left(\sum_{i=1}^{N} x_{s_i} y_{s_i}\right) = \left(\sum_{g=1}^{G} x_g' x_g\right)^{-1} \left(\sum_{g=1}^{G} x_g' y_g\right)$$

using individual-level and group-level notation respectively.

The most common approach to inference with weakly dependent data is to use a 'plug-in' estimator, call it $\widetilde{V}_N$, of the variance matrix of $x_{s_i}\varepsilon_{s_i}$, along with the usual large-sample approximation for the distribution of $\widehat{\beta}_N$. Specifically, the large-sample distribution of $\hat{\beta}_N$ is

$$\sqrt{N}\left(\hat{\beta}_N - \beta\right) \xrightarrow{d} N(0, Q^{-1}VQ^{-1})$$
$$V = \lim_{N\to\infty} Var(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_{s_i}\varepsilon_{s_i})$$

where Q is the limit of the second moment matrix for x. The typical method uses the sample average of $x_{s_i}x_{s_i}^{'}$ to estimate Q and plugs in a consistent estimator, $\widetilde{V}_N$, of V to arrive at the approximation:

$$\hat{\beta}_N \overset{Approx}{\sim} N\left(\beta, \frac{1}{N}\left[\frac{1}{N}\sum_{i=1}^{N} x_{s_i}x_{s_i}^{'}\right]^{-1} \widetilde{V}_N \left[\frac{1}{N}\sum_{i=1}^{N} x_{s_i}x_{s_i}^{'}\right]^{-1}\right) \tag{2.1}$$

Conventionally, one would use an estimator $\widetilde{V}_N$ that is consistent for V under general forms of heteroskedasticity and autocorrelation in $x_{s_i}\varepsilon_{s_i}$. Such estimators are commonly referred to as HAC variance estimators; see, for example, Newey and West (1987), Andrews (1991), and Conley (1999). In the remainder, we refer to HAC estimators as $\widehat{V}_{HAC}$.

When the data is located at integers on the line, say $s_1 = 1, ..., s_N = N$, spatial and discrete time series estimators for V coincide and typically are written as a weighted sum of sample autocovariances with weights, $W_N(\cdot)$, depending on the lag/gap between observations:

$$\hat{V}_{HAC} = \sum_{h=-(N-1)}^{N-1} W_N(h) \frac{1}{N} \sum_j x_{s_j} e_{s_j} x'_{s_{j+h}} e_{s_j+h}$$

where e_{s_j} in this expression is an OLS residual. This estimator will be consistent under regularity conditions that include an assumption that $W_N(h) \to 1$ for all h slowly enough for the variance of $\hat{V}_{HAC}$ to vanish as $N \to \infty$; see, e.g., Andrews (1991). Perhaps the most popular choice for weight function $W_N(h)$ is the Bartlett kernel: an isoceles triangle that is one at $h = 0$ with a base of width $2H_N$: $W_N(h) = (1 - \frac{|h|}{H_N})^+$.

To see the link between $\hat{V}_{HAC}$ above and HAC estimators in other metric spaces, it is useful to rewrite $\hat{V}_{HAC}$ using "row and column" notation to enumerate all pairs of cross products rather than organizing them by lag/gap. The above expression for $\hat{V}_{HAC}$ can be written as

$$\hat{V}_{HAC} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} W_N(s_i - s_j) x_{s_i} e_{s_i} x'_{s_j} e_{s_j}.$$

Thus $\hat{V}_{HAC}$ is a weighted sum of all possible cross products of $x_{s_i} e_{s_i}$ and $x'_{s_j} e_{s_j}$. The weights depend on the lag/gap between the observations, i.e. their distance. This idea generalizes immediately to higher dimensions (and other metric spaces) yielding a HAC estimator:

$$\hat{V}_{HAC} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} W_N(dist(s_i, s_j)) x_{s_i} e_{s_i} x'_{s_j} e_{s_j}$$

where $dist(s_i, s_j)$ gives the distance between observations located at s_i and s_j. Regularity conditions for this estimator are analogous to those for locations on the line. Key among these conditions is

that $W_N(d) \to 1$ for all d slowly enough for the variance of $\hat{V}_{HAC}$ to vanish as $N \to \infty$; see Conley (1999). The typical empirical approach is to choose a weight function $W_N(\cdot)$ and compute $\widehat{V}_{HAC}$ to plug into expression (2.1).

In a time series setting, Kiefer and Vogelsang (2002, 2005) (KV) provide an alternative way to conduct inference using HAC estimators. They focus on $\hat{V}_{HAC}$ defined with an H_N sequence that violates the conditions for consistency. In particular, H_N grows at the same rate as the sample size, and thus $\widehat{V}_{HAC}$ converges to a non-degenerate random variable. They then calculate the large-sample distribution for usual test statistics formed with this random-variable-limit $\hat{V}_{HAC}$ matrix. The resulting limit distributions for test statistics are non-standard. However, they turn out to not depend on parameters of the data generating process (i.e., they are pivotal), so critical values can be tabulated via simulation. KV provide convincing evidence that inference based on this approximation outperforms the plug-in approach in the time series context. Bester, Conley, Hansen, and Vogelsang (2008) provide similar results in a spatial context.

2.2. **Our Approach**

Our main approach in this paper is in the spirit of KV. We use an asymptotic sequence in which the estimator of V, the cluster covariance estimator (CCE), is not consistent but converges in distribution to a limiting random variable. The CCE is computationally very tractable and is already familiar to many applied researchers. The inference procedure we propose is therefore easy to implement and remains valid when the data are indexed in high-dimensional spaces (e.g., a panel or a cross section with dependence along multiple dimensions). The CCE may be defined as follows:

$$\hat{V}_N \equiv \frac{1}{N} \sum_{g=1}^{G_N} x_g' e_g e_g' x_g$$

using group notation. The same estimator can of course also be written using individual observation notation as

$$\hat{V}_N = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} 1(i,j \in \textit{same group}) x_{s_i} e_{s_i} x_{s_j}^{'} e_{s_j}.$$

Thus $\hat{V}_N$ can be thought of as a HAC estimator with a nonstandard weighting kernel. Instead of weights that depend on distances between observations, it has a uniform weight function that indicates common group membership.

The CCE is commonly employed along with an assumption of independence across groups; see, e.g., Liang and Zeger (1986), Arellano (1987), Wooldridge (2003), Bertrand, Duflo, and Mullainathan (2004), and Hansen (2007). It is important to note that we are *not* assuming such independence. Instead we assume our data are weakly dependent with dependence structure described by observations' indices. The results in this paper cover cases where $G_N = G$ is taken as fixed, so that $\hat{V}_N$ converges to a random variable and thus is not a consistent estimator of V.[5]

Our method uses $\hat{V}_N$ to form an estimator of the asymptotic variance of $\widehat{\beta}$ given in equation (2.1) and then uses this estimate of the asymptotic variance to form usual t and Wald statistics. We calculate limiting distributions for these t and Wald statistics under a sequence that holds G fixed as $L_N \to \infty$. Under a general set of assumptions, the limiting distribution of the t-statistic is $\sqrt{\frac{G}{G-1}}$ times a Student-t distribution with $G-1$ degrees of freedom, and a Wald statistic with q restrictions has a limiting distribution that is $\frac{Gq}{G-q}$ times an $F_{q,G-q}$ distribution. Confidence sets can be obtained using these distributions in the usual fashion. The CCE is also trivial to estimate with most standard econometrics packages. For example, the t-statistics created via the cluster command in Stata 10 can be directly used to implement our inference method if they are used with critical values of a Student-t with G-1 degrees of freedom.[6]

[5]See Bester, Conley, and Hansen (2010) for a demonstration that $\hat{V}_N$ is consistent for V if G_N and L_N grow at appropriate rates.

[6]The exact scaling in Stata 10 is slightly different than ours due to the presence of a small-sample degrees of freedom correction. Specifically, $\widehat{V}_{STATA} = \frac{N-1}{N-k}\frac{G}{G-1}\widehat{V}_N$; see *Stata User's Guide Release 10* p. 275.

Throughout the paper we will refer to a partitioning of the data into groups of observations defined by the researcher. The idea is to construct large groups that are shaped properly for within-group averages/sums to be approximately independent. As suggested by our simulation results, this will often require the number of groups be kept small. We consider equal-sized groups corresponding to contiguous locations.[7] In m-dimensions, we impose the additional restriction that the size of group boundaries relative to their volume is the same order as for m-dimensional cubes. The contiguity and boundary conditions imply that, in large groups, most of the observations will be interior and far from points in other groups. Under weak dependence, these interior points will then be approximately independent across groups. Therefore, the set of near-boundary points will be sufficiently limited for their influence upon correlations across group-level averages to be vanishingly small.

3. **Asymptotic Properties**

In this section, we develop the asymptotic properties of the CCE with weakly dependent data. We first state results under an asymptotic sequence, which we refer to as "fixed-G", that takes the number of groups as fixed and lets the number of observations in each group become arbitrarily large. Under this sequence, we show that the CCE is not consistent but converges in distribution to a limiting random variable. We show that, under sensible sampling conditions, standard t and Wald statistics formed using the CCE follow limiting t and F distributions.

Thus, scaling the STATA t-statistic by multiplying it by $\sqrt{\frac{N-1}{N-k}}$ would be equivalent to our recommended procedure. There is unlikely to be any appreciable difference between using this reweighting and directly using the reported cluster t-statistics since $\frac{N-1}{N-k}$ will be close to one in many applications. Also, since $\frac{N-1}{N-k}$ will always be greater than one, using the statistic from STATA without modification will in a sense be conservative. We note that the confidence intervals reported by Stata after the use of the cluster command use critical values from a Student-t distribution with $G-1$ degrees of freedom.

[7] I.e. groups do not have gaps.

3.1. **Fixed-G Asymptotics**

We provide a simple set of conventional regularity conditions that are sufficient to obtain our fixed-G results. Assumption 1 contains a set of mixing and moment conditions and Assumption 2 contains a set of restrictions upon the nature of groups. These two assumptions yield Lemma 1 which shows that a central limit theorem applies within each group and groups are asymptotically uncorrelated. Note that Lemma 1 is the central ingredient to the theoretical results and will be implied by a variety of other sets of conditions.

For simplicity, we will index observations on an m-dimensional integer lattice, $\mathbb{Z}^m$, and use the maximum coordinatewise metric $dist(s_i, s_j)$.[8] Throughout, let $\mathcal{G}_{g_1}$, $\mathcal{G}_{g_2}$ be two disjoint sampling regions (index sets) corresponding to groups $\{g_1, g_2\} \subseteq \{1, \ldots, G\}$ with $g_1 \neq g_2$. Use $|\mathcal{G}|$ to refer to the number of elements in the region. The boundary of a region is defined as $\partial\mathcal{G} = \{i \in \mathcal{G} : \exists j \notin \mathcal{G} \text{ s.t. } dist(s_i, s_j) = 1\}$. We now state sufficient conditions for our main results in the form of Assumptions 1 and 2:

Assumption 1.

(i) The sample region grows uniformly in m non-opposing directions as $N \to \infty$.

(ii) As $N \to \infty$, $L_N \to \infty$ and G is fixed.

(iii) $\{x_s, \varepsilon_s\}$ is α-mixing[9] and satisfies (a) $\sum_{j=1}^{\infty} j^{m-1} \alpha_{1,1}(j)^{\delta/(2+\delta)} < \infty$, (b) $\sum_{j=1}^{\infty} j^{m-1} \alpha_{k,l}(j) < \infty$ for $k + l \leq 4$, and (c) $\alpha_{1,\infty}(j) = O(j^{-m-\eta})$ for some $\delta > 0$ and some $\eta > 0$.

[8]The maximum coordinatewise distance metric is defined as $dist(s_i, s_j) = \max_{l \in \{1,\ldots,m\}} |s_i(l) - s_j(l)|$ where $s_i(l)$ is the l^{th} element of vector s_i. Note that for $s_i \neq s_j$, $dist(s_i, s_j)$ takes values in the positive integers. Note that Lemma 1 may be established with observations belonging to different spaces and using different metrics.

[9]We use the standard notion of an α- or strong mixing process from time series. See, for example, White (2001) Definition 3.42. For spatial processes, we use a mixing coefficient for a random field defined as follows. Let $\mathcal{F}_\Lambda$ be the σ-algebra generate by a given random field ψ_{s_m}, $s_m \in \Lambda$ with Λ compact, and let $|\Lambda|$ be the number of $s_m \in \Lambda$. Let $\Upsilon(\Lambda_1, \Lambda_2)$ denote the minimum distance from an element of Λ_1 to an element of

$\sup_s \mathrm{E}|\varepsilon_s|^{2r} < \infty$ and $\sup_s \mathrm{E}|x_{sh}|^{2r} < \infty$ for $r > 2+\delta$ where x_{sh} is the h^{th} element of vector x_s. $\mathrm{E}[\frac{1}{L_N} x_g' x_g]$ is uniformly positive definite with constant limit Q_g for all $g = 1, ..., G$.

(iv) $\mathrm{E}[x_s \varepsilon_s] = 0$. $V_{Ng} = \mathrm{var}[\frac{1}{\sqrt{L_N}} x_g' \varepsilon_g]$ is uniformly positive definite with constant limit Ω_g for all $g = 1, ..., G$.

Part (i) ensures that indexing in m-dimensions is required, i.e. that indexing in a lower dimension space is not adequate to describe the dependence in the data. Part (ii) restates the asymptotic sequence which has a fixed number of groups whose size is increasing. The key part of (iii) is the mixing and moment conditions. The conditions allow for quite general forms of heteroskedasticity and non-stationarity, though we will restrict these further to establish the key results.

Assumption 2 (Restrictions on groups)**.**

(i) *Groups are mutually exclusive and exhaustive.*

(ii) *For all g,* $|\mathcal{G}_g| = L_N$.

(iii) *Groups are contiguous in the metric dist*$(\cdot)$.

(iv) *For all g,* $|\partial\mathcal{G}| < C L_N^{\frac{m-1}{m}}$.

Part (i) of Assumption 2 could be relaxed to allow an asymptotically negligible amount of overlap across groups or omission of an asymptotically negligible portion of the data. Part (ii) of Assumption 2 assumes a common group size.[10] Part (iii) of Assumption 2 simply requires that groups are connected but could be relaxed to allow a finite number of disjoint components for a group. Assumption 2(iii)-(iv) imply that asymptotically groups correspond to regions of the sampling space

Λ_2. For our results, we use the maximum coordinate-wise distance metric. The mixing coefficient is then $\alpha_{k,l}(j) \equiv \sup\{|P(A \bigcap B) - P(A)P(B)|\},\ A \in \mathcal{F}_{\Lambda_1},\ B \in \mathcal{F}_{\Lambda_2}$, and $|\Lambda_1| \leq k$, $|\Lambda_2| \leq l$, $\Upsilon(\Lambda_1, \Lambda_2) \geq j$. Mixing requires that $\alpha_{k,l}(j)$ convereges to zero as $j \to \infty$.

[10]We ignore integer problems for notational convenience and simplicity. If we allowed different group sizes, say L_g, all results would carry through immediately as long as $L_{g_1}/L_{g_2} \to 1$ for all g_1 and g_2. We provide results for when group sizes are not asymptotically equivalent in the appendix.

that resemble a collection of regular polyhedra growing to cover the space. In the special case of a time series ($m = 1$), (iii)-(iv) requires that groups are 'blocks' on the line and, for example, rules out groups consisting of every kth observation. In general, these conditions ensure that the majority of points within each group are in the group's interior as groups get larger. The boundary condition in (iv) is the key element used for our results.

Lemma 1. *Under Assumptions 1 and 2 as $L_N \to \infty$,*

$$(i) \qquad \frac{1}{L_N}\begin{pmatrix} x_1'x_1 \\ \vdots \\ x_G'x_G \end{pmatrix} \xrightarrow{p} \begin{pmatrix} Q_1 \\ \vdots \\ Q_G \end{pmatrix} \text{ and}$$

$$(ii) \qquad \frac{1}{\sqrt{L_N}}\begin{pmatrix} x_1'\varepsilon_1 \\ \vdots \\ x_G'\varepsilon_G \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} \Omega_1 & & 0 \\ & \ddots & \\ 0 & & \Omega_G \end{pmatrix}\right)$$

for Q_g and Ω_g positive definite for all $g = 1, ..., G$.

Lemma 1 states that a suitable law of large numbers applies to $L_N^{-1}x_g'x_g$ and that $L_N^{-\frac{1}{2}}x_g'\varepsilon_g$ obeys a central limit theorem with zero asymptotic covariance across groups. The dependence restrictions in Assumption 1 are sufficient to verify that a central limit theorem applies to the $L_N^{-\frac{1}{2}}x_g'\varepsilon_g$. As usual with clustering estimators, no assumptions are made about the structure of Q_g or Ω_g beyond their being positive definite. We note again that, unlike other treatments of clustering estimators, groups need not be independent for any finite group size L_N.

When G is fixed and $L_N \to \infty$, Lemma 1 is sufficient to characterize the behavior of the CCE. In this case, $\hat{V}_N$ is not consistent, but converges to a limiting random variable. In general, the reference distributions for test statistics based on the CCE are not pivotal and are nonstandard under this sequence. However, we also consider two mild forms of homogeneity under which reference

distributions for the usual t and Wald statistics simplify to the usual t- and F-distributions with degrees of freedom determined by the number of groups.

Assumption 3 (Homogeneity of $x_g'x_g$)**.** *For all g,* $Q_g \equiv Q$.

Assumption 4 (Homogeneity of $x_g'\varepsilon_g$)**.** *For all g,* $\Omega_g \equiv \Omega$.

Assumptions 3 and 4 respectively assume that the design matrices $x_g'x_g$ converge to the same limit within each group and that the asymptotic variances of the within-group scores are the same across groups. These conditions are implied by covariance stationarity of the individual observations but may also be satisfied even if covariance stationarity is violated. It is interesting to consider these conditions in a time series context. Functional central limit theorems (FCLTs) are commonly employed in providing asymptotic results for time series estimators and inference procedures. In the time series case, any grouping with approximately equal numbers of adjacent observations within each block[11] will satisfy Assumption 2. An FCLT will also imply Assumptions 1, 3, and 4. Thus, Assumptions 1-4 are applicable in any application where one believes an FCLT applies. It is also clear that Assumptions 1, 3, and 4 will be satisfied in many cases where an FCLT would not apply as they allow for substantial within group heterogeneity.

We are now ready to state our main results. Let $\hat{Q} = \frac{1}{N}\sum_g x_g'x_g$. In the following, consider testing $H_0 : R\beta = r$ against $H_1 : R\beta \neq r$ where R is $q \times k$ and r is a q-vector using the test statistics

$$\hat{F} = N\left(R\widehat{\beta} - r\right)'\left[R\hat{Q}^{-1}\hat{V}_N\hat{Q}^{-1}R'\right]^{-1}\left(R\widehat{\beta} - r\right),$$

or, when $q = 1$,

$$\hat{t} = \frac{\sqrt{N}\left(R\widehat{\beta} - r\right)}{\sqrt{R\hat{Q}^{-1}\hat{V}_N\hat{Q}^{-1}R'}}.$$

Properties of $\hat{t}$ and $\hat{F}$ are given in the following proposition.

[11] For example, group one has observations from years 1,...,10; group 2 has observations from years 11,...20; etc.

Proposition 1. *Suppose* $\{\mathcal{G}_g\}$ *is defined such that* $L_N \to \infty$ *and* G *is fixed as* $N \to \infty$ *and that Lemma 1 holds. Let* $B_g \sim N(0, I_k)$ *denote a random* k*-vector and* $\Omega_g = \Lambda_g \Lambda_g'$*. Define matrices* $\mathbf{Q}$ *and* $\mathbf{S}$ *such that* $\mathbf{Q} = \sum_g Q_g$ *and* $\mathbf{S} = \sum_g \Lambda_g B_g$*. Then,*

i. $\hat{V}_N \xrightarrow{d} V_A = \frac{1}{G}\sum_g \left[\Lambda_g B_g B_g' \Lambda_g' - Q_g \mathbf{Q}^{-1}\mathbf{S} B_g' \Lambda_g' - \Lambda_g B_g \mathbf{S}'\mathbf{Q}^{-1} Q_g + Q_g \mathbf{Q}^{-1}\mathbf{S}\mathbf{S}'\mathbf{Q}^{-1} Q_g\right]$*, and under* H_0*,*

$$\hat{t} \xrightarrow{d} \frac{\sqrt{G} R\mathbf{Q}^{-1}\mathbf{S}}{\sqrt{R(\mathbf{Q}/G)^{-1} V_A (\mathbf{Q}/G)^{-1} R'}} \qquad \text{and}$$
$$\hat{F} \xrightarrow{d} G\mathbf{S}'\mathbf{Q}^{-1}R' \left[R(\mathbf{Q}/G)^{-1} V_A (\mathbf{Q}/G)^{-1} R'\right]^{-1} R\mathbf{Q}^{-1}\mathbf{S}.$$

ii. *if Assumption 3 is also satisfied,* $\hat{t} \xrightarrow{d} \sqrt{\frac{G}{G-1}} t^*_{G-1}$ *under* H_0 *where* t^*_{G-1} *satisfies*

$$P\left(|t^*_{G-1}| > c_{G-1}(\alpha)\right) \leq \alpha$$

for $c_{G-1}(\alpha)$ *the usual critical value for an* $\alpha-$*level two-sided t-test based on a t-distribution with* $G-1$ *degrees of freedom for any* $\alpha \leq 2\Phi(-\sqrt{3})$ *and for any* $\alpha \leq 0.1$ *if* $2 \leq G \leq 14$*.*

iii. *if Assumptions 3 and 4 are also satisfied,* $\hat{t} \xrightarrow{d} \sqrt{\frac{G}{G-1}} t_{G-1}$ *and* $\hat{F} \xrightarrow{d} \frac{Gq}{G-q} F_{q,G-q}$ *under* H_0 *where* t_{G-1} *and* $F_{q,G-q}$ *are respectively random variables that follow a* t *distribution with* $G-1$ *degrees of freedom and an* F *distribution with* q *numerator and* $G-q$ *denominator degrees of freedom.*

The results of Proposition 1 are stated under increasingly more restrictive homogeneity assumptions. The benefit of additional homogeneity is that the limiting behavior of test statistics is determined by standard t- and F- distributions. Using these standard reference distributions makes performing hypothesis tests and constructing confidence intervals as easy as under the normal asymptotic approximations, and we show in simulation examples that these approximations perform well in finite samples. The results also clearly illustrate the intuition that the behavior of test statistics under weak dependence is essentially governed by the number of 'approximately uncorrelated observations' in the sample, which in this case corresponds to the number of groups.

Proposition 1, part (i) does not impose homogeneity and implicitly allows for group sizes that are not asymptotically equivalent. Without further restrictions, usual test statistics formed using the CCE converge in distribution to a ratio of random variables. These limiting distributions are neither standard nor pivotal though one could attempt to estimate the nuisance parameters involved in the distributions and simulate from them to conduct inference.

We note that, under sequences where $G_N \to \infty$, the reference distributions obtained in Parts (ii) and (iii) of Proposition 1 are still valid in the sense that they converge to the usual normal and χ^2 reference distributions as $G_N \to \infty$.[12] That the approximate distributions obtained in Parts (ii) and (iii) of Proposition 1 will remain valid in either asymptotic sequence, while the usual normal and χ^2 approximations will only be valid under sequences when G_N is arbitrarily large, strongly suggests that one should always simply use the fixed-G limits. Simulation results reported in Section 4 provide strong support for this conclusion.

The result in Part (ii) of Proposition 1 shows that under a homogeneity assumption on the limiting behavior of the design matrix across groups, the usual t-statistic converges to $\sqrt{G/(G-1)}$ times a random variable with tail behavior similar to a t_{G-1} random variable, where by similar we mean that the test will reject with probability less than or equal to the nominal size of a test as long as the test is at a small enough level of significance (less than around .08 in general). This result suggests that valid inference may be conducted by simply rescaling the usual t-statistic by $\sqrt{(G-1)/G}$ which is equivalent to using $\frac{G}{G-1}\hat{V}_N$ as the covariance matrix estimator. This result uses Theorem 1 of Ibragimov and Müller (2006); see also Bakirov and Székely (2005). To our knowledge, there is no currently available similar result for $\hat{F}$.

[12] With additional technical conditions, it can be shown that Proposition 1 part (i) implies that the usual normal and χ^2 reference distributions will be valid under a sequential asymptotics where first $L_N \to \infty$ and then $G_N \to \infty$. We do not pursue this since this sequence is not immediately useful when G is fixed and provides a similar result to that obtained in the working paper Bester, Conley, and Hansen (2010) under asymptotics where $\{L_N, G_N\} \to \infty$ jointly.

The final results in part (iii) show that under a homogeneity assumption on the limiting behavior of the design matrices and on the within-group asymptotic variance, the usual t- and Wald statistics converge to scaled t- and F-distributions.[13] The scale on the t-statistic is again $\sqrt{G/(G-1)}$ which suggests using $\frac{G}{G-1}\hat{V}_N$ as the covariance matrix estimator if one is interested in inference about scalar parameters or rank one tests. On the other hand, the scale of the F-distribution depends on the number of parameters being tested, though rescaling the F-statistic appropriately is trivial.

Overall, the conclusions of Proposition 1 are useful from a number of standpoints. The asymptotic distributions provided in Parts (ii) and (iii) of Proposition 1 are easier to work with than KV distributions on the line, and this difference becomes more pronounced in higher dimensions. Our approximations should also more accurately account for the uncertainty introduced due to estimating the covariance matrix than plug-in approaches. This improvement is evidenced in a simulation study reported below where we find that using the reference distributions implied by the fixed-G asymptotic results eliminates a substantial portion of the size distortion that occurs when using HAC estimators plugged into a limiting normal approximation.

Our fixed-G results are related to inference results obtained for time series HAC in KV and spatial HAC in Bester, Conley, Hansen, and Vogelsang (2008) (BCHV). BCHV provide 'fixed-b' results analogous to KV in the context of inference based on spatial HAC estimators. Relative to KV and BCHV, we use the CCE in place of a HAC estimator which allows us to obtain simple, pivotal limiting behavior for test statistics under weaker homogeneity assumptions than those employed in KV and BCHV. Specifically, KV and BCHV make use of time series and spatial FCLTs that produce pivotal but nonstandard reference distributions for test statistics. In our context, satisfaction of these FCLTs would impose within group homogeneity in addition to across group homogeneity and place further restrictions on the regularity of the sample region. The spatial FCLT will also generally need a specified metric for distances between observations and require correct measures

[13] We thank Jim Stock and Mark Watson for pointing out the F-statistic result. See also Stock and Watson (2008).

of each observations location. Assumptions 1, 3, and 4 are guaranteed by the FCLTs in KV and BCHV but are not strong enough to ensure that these FCLTs apply.

It is important to note the relationship between our fixed-G approach and the Fama-Macbeth type estimator of Ibragimov and Müller (2006) (IM) mentioned above. The IM approach is to partition the data into groups and separately estimate the model parameters using each group. Inference for a scalar parameter is then conducted using a t-statistic with the simple average of the group-level estimates in the numerator and the standard deviation of the estimates across groups in the denominator.[14] Under our Lemma 1, IM show that inference based on these t-statistics, using critical values from a t-distribution (with degrees of freedom one less than the number of groups) is asymptotically valid.

Our approach and that of IM both rely on the conclusion of Lemma 1, that group averages are asymptotically Gaussian and independent. In IM this is a high-level assumption. They do provide primitive conditions sufficient for the conclusion of Lemma 1 to hold in the case where the groups consist of consecutive observations of weakly dependent data on the line (e.g., time series). Our paper provides a set of primitive conditions that are sufficient for Lemma 1 to hold in sampling environments where dependence is indexed in m-dimensions.

Since both papers rely on Lemma 1, the primitive conditions in Assumptions 1 and 2 imply that the IM results are also valid with empirically relevant forms of m-dimensionally indexed dependence. Interestingly, the result obtained in IM only makes use of Lemma 1, so their t-statistic based result remains valid when Assumption 4 is not satisfied. This is particularly useful in applications with pronounced heterogeneity in regressor variances across groups where our Assumption 4 would not apply but the IM approach remains valid.

[14] Our approach differs in that the numerator is based on a parameter estimate using data from the entire sample, and the CCE is used as the denominator.

Our approach and that of IM are best thought of as complements as there are clearly scenarios where each would be preferred over the other.[15] For tests of scalar hypotheses, our proposed t-statistic and the t-statistic of IM differ in both their numerator and denominator, which complicates a general comparison of the two approaches. However, we can say something about scenarios where we anticipate these tests' performance will differ. For example, when there is substantial finite sample bias, due e.g. to instrumental variables, our approach may perform better because the numerator uses a point estimator based on the full sample, rather than an average of group-level estimators whose finite sample biases will generally not average out. The IM approach should outperform ours when group-level point estimators have minimal bias and pronounced heterogeneity in variances. We provide simulation results in Section 4.3 to illustrate the relative performance of our methods and those of IM across scenarios with differing magnitudes of bias and variance heterogeneity. Both approaches are simple to implement in practice and offer substantial improvements relative to existing inference methods with dependent data, and both should therefore prove useful to applied researchers.

4. **Simulation Examples**

The previous section provides a limiting result under an asymptotic sequence when the number of groups remains fixed. Under this sequence, standard test-statistics follow asymptotic t- or $F-$

[15]IM is in show that simple (unweighted) averages of estimates obtained in subsambles and the standard deviation of these estimates can be used to perform heteroskedasticity and autocorrelation robust inference about a scalar parameter. There is currently no known robustness result for joint inference. Of course, one may conduct joint hypothesis tests and inference in the Fama-MacBeth-IM framework by estimating the variances of the within group estimators and appropriately weighting. IM explicitly try to avoid this. Under homogeneity, they would obtain similar results to ours under homogeneity and would need to estimate the same quantities we would need to estimate if we were to use the limiting distribution given in Proposition 1.i more generally.

distributions, which are extremely easy to use and should work better in finite samples than the usual asymptotic approximations. In this section, we provide evidence on the inference properties of tests based on the CCE, first using simulation experiments in entirely simulated data, and then for experiments in which we regress actual unemployment rates on simulated treatments. The latter experiments are conducted in a panel data setting where time and state-level fixed effects are included. In these simulations, we consider inference about a slope coefficient from a linear regression model with point estimates obtained using OLS. We also conduct a separate set of simulation results with the explicit goal of comparing our approach with the one proposed in Ibragimov and Müller (2006). We use the 2SLS estimator in these simulations as we are interested in the effects of biases in the numerator of our test statistics as well as heterogeneity in regressor variances.

4.1. **Results using Simulated Treatments and Outcomes**

We consider two basic types of DGP: an autoregressive time series model and a low-order moving average spatial model. For both models, we set

$$y_s = \alpha + x_s\beta + \varepsilon_s,$$

where x_s is a scalar, $\alpha = 0$, and $\beta = 1$. For the time series specification, we generate x_s and ε_s as

$$x_s = 1 + \rho x_{s-1} + v_s, \quad v_s \sim N(0,1) \text{ and}$$

$$\varepsilon_s = \rho\varepsilon_{s-1} + u_s, \quad u_s \sim N(0,1)$$

with initial observation generated from the stationary distribution of the process. We consider three different values of ρ, $\rho \in \{0, .5, .8\}$ and set $N = 100$.

In the spatial case, we consider data generated on a $K \times K$ integer lattice. We generate x_s and ε_s as

$$x_s = \sum_{\|h\| \leq 2} \gamma^{\|h\|} v_{s+h},$$

$$\varepsilon_s = \sum_{\|h\| \leq 2} \gamma^{\|h\|} u_{s+h}$$

with $\|h\| = \text{dist}(0, h)$ in this expression, $u_s \sim N(0,1)$, and $v_s \sim N(0,1)$ for all i and j. We consider three different values of γ, $\gamma \in \{0, .3, .6\}$ and set $K = 36$ for a total sample size of $N = 1296$.[16]

Table 1 reports rejection rates for 5% level tests from a Monte Carlo simulation experiment. The time series simulations are based on 30,000 simulation replications and the spatial simulations are based on 500 simulation replications. Row labels indicate which covariance matrix estimator is used. Column 2 indicates which reference distribution is used with KV corresponding to the Kiefer and Vogelsang (2005) approximation. Rows labeled IID and Heteroskedasticity use conventional OLS standard errors and heteroskedasticity robust standard errors respectively. Rows labeled Bartlett use HAC estimators with a Bartlett kernel. Rows labeled CCE use the CCE estimator. For tests based on IID and Heteroskedasticity, a N(0,1) distribution is used as the reference distribution. For the CCE estimator, a t(G-1) distribution is used as the reference distribution. For the HAC estimator, we consider two different reference distributions: a N(0,1) and the Kiefer and Vogelsang (2005) approximation. Small, Medium, and Large denote lag truncation parameters for HAC or number of observations per group for CCE. For time series models, Small, Medium, and Large respectively denote lag truncation at 12, 20, and 38 for HAC and denote numbers of groups of 12, 8, and 4 for CCE . For spatial models, Small, Medium, and Large denote lag truncation at 14, 122, and 486 for HAC and denote numbers of groups of 144, 16, and 4 for CCE.[17]

Looking first at the time series results, we see that tests based on the CCE with a small number of groups perform quite well across all of the ρ parameters considered. As expected, the tests based

[16]We draw u_s and v_s on a 40×40 lattice to generate the 36×36 lattice of x_s and ε_s.

[17]We chose these truncation parameters for the Bartlett kernels by taking $(3N)/(2G)$ where N is the sample size in the simulation and G is the number of groups used in the CCE. This rule of thumb produces HAC estimators with roughly the same variance as the corresponding CCE estimator. We thank a referee for pointing this out.

on the CCE overreject with ρ of .8 when a moderate or large number of groups is used, though size distortions are modest with ρ of .5 for all numbers of groups. Comparing across HAC and the CCE, we see that tests based on the HAC estimator using the usual asymptotic approximation have large size distortions. Looking at the HAC rejection frequencies closest to the nominal level of 5%, we see that the HAC tests reject 11.6% of the time with $\rho = .5$ and 17.7% of the time with $\rho = .8$ compared to 6.0% of the time and 8.2% of the time for the CCE-based tests. Tests based on the Kiefer and Vogelsang (2005) approximation behave similarly to tests based on the CCE, highlighting the similarity between the fixed-G approach for the CCE and the "fixed-b" approach for HAC estimators. The results also demonstrate the well-known result that conducting inference without accounting for serial correlation leads to tests with large size distortions.

The spatial results follow roughly the same pattern as the time series results. Tests based on the CCE with a small number of groups perform uniformly quite well regardless of the strength of the correlation. In the moderate and no correlation cases, we also see that the CCE-based tests do reasonably well when more groups are used.

Size-adjusted power curves comparing tests using HAC with the KV reference distribution to CCE are fairly similar across the designs considered. We report the case with the largest discrepancy between power curves in Figure 1.[18] Figure 1 provides power curves for the test based on the CCE with four groups (the solid curve) and the HAC estimator (the curve with x's) with a smoothing parameter of 38 using the Kiefer and Vogelsang (2005) reference distribution for the time series case with $\rho = 0.8$. We can see that there is a modest power loss due to using tests based on the CCE relative to HAC with Bartlett kernel with smoothing parameters that produce similar size in this figure. We note that the power loss is much smaller across the remaining designs. We note that this power loss of the CCE is accompanied by slightly smaller size distortions relative to the conventional HAC procedure with KV critical values.

[18]We choose to focus on power for procedures with approximately correct and comparable size.

4.2. **Results using Unemployment Rate Outcomes**

In our second set of simulations, we use the log of monthly state-level unemployment rates as our dependent variable.[19] The data we consider have monthly unemployment rates for each state from 1976 to 2007 giving a total of 384 months in the sample. We discard Alaska and Hawaii but include Washington D.C. giving us 49 cross-sectional observations. We regress these unemployment rates on a randomly generated treatment. These simulations allow us to examine the properties of CCE-based inference using our fixed-G approximations for data with a strong spatial and inter-temporal correlation structure determined by actual unemployment outcomes. In these simulations, we only consider clustering based methods as these are the most commonly employed methods used to do inference in applied microeconomics with panel data.

In this section, we consider inference on the slope coefficient from the model

$$\log(y_{st}) = \beta x_{st} + \alpha_s + \alpha_t + \varepsilon_{st}$$

where y_{st} is the unemployment rate in state s at time t, α_s and α_t are respectively unobserved state and time effects, ε_{st} is the error term, and x_{st} is a simulated treatment whose generation we discuss below. In all of the simulations, we set $\beta = 0$ and treat α_s and α_t as fixed effects.[20] In most simulations, we first-difference to remove α_s and include a full set of time dummies, though we include one set of results where we estimate β including a full set of both state and month

[19] We use seasonally unadjusted monthly state-level unemployment rates from the BLS available at ftp://ftp.bls.gov/pub/time.series/la/.

[20] We note that one still estimates appreciable spatial and intertemporal correlation in log unemployment rates even after accounting for state and time effects. This is also apparent in the simulation results as the scores would be uncorrelated if there were no spatial or serial correlation in log unemployment after accounting for state and time effects.

dummies.[21] We note that this is a simple but fairly standard specification in applied research and that it is similar to models considered in Shimer (2001) and Foote (2007).

We generate a treatment that is meant to represent a variable such as the log of the youth employment share as considered in Shimer (2001) and Foote (2007). We generate the treatment to be both spatially and intertemporally correlated from the model

$$
\begin{aligned}
x_{st} &= \sigma\left(u_{st} + \gamma \sum_{d(s,r)=1} u_{rt}\right) \text{ where} \\
u_{st} &= \sum_{j=1}^{p} \rho_j u_{s(t-j)} + v_{st}, \\
v_{st} &\sim N(0,1),
\end{aligned}
$$

$d(s,r)$ is one for adjacent states s and r and zero otherwise, and $\rho_1, ..., \rho_p$ and γ respectively control the intertemporal and spatial dynamics. We consider an AR(13) with $\rho_1 = .95$, $\rho_{12} = .68$, and

[21]The inclusion of fixed effects can complicate the analysis. When groups defined by fixed effects do not cross the user-defined boundaries used to define clusters for inference with the CCE, the results of Proposition 1 carry through immediately for inference about common parameters. However, there is an important interaction between eliminating both state and time fixed effects and grouping in the panel context. When N is large and T is fixed, removing both state and time effects via demeaning (equivalently using a full set of state and time dummies) is fine as long as groups are formed by splitting individuals into different groups that include all time series observations for the same individual in the same group. Such a strategy is what intuition would suggest since a large N small T case is equivalent to a vector cross-section. Similarly, in large T small N settings, inference should be regarded as for a vector time series and groups formed by including all individuals and splitting over time. In large N large T cases, the inclusion of both state and time dummies alters the fixed-G reference distribution from that presented in Proposition 1. However, the result in Proposition 1 remains valid when individual effects are removed by first-differencing and time dummies are included.

$\rho_{13} = -.68$ which was chosen because it matches the empirical intertemporal dynamics in $\log(y_{st})$. We also consider an AR(1) with $\rho_1 = .8$ as a simple benchmark.[22]

We report simulation results in Table 2. In all cases, we report rejection frequencies for 5% level tests of the hypothesis that $\beta = 0$. Rows labeled IID and Heteroskedasticity use conventional OLS and heteroskedasticity consistent standard errors respectively. The remaining rows use the CCE with different grouping schemes. "State" and "Month" use states and months as groups, respectively. "State/Month" treats observations as belonging to the same group if they belong to the same state or the same month; the variance matrix for this metric can be estimated by summing the CCE with groups defined by states and the CCE for groups defined by months and then subtracting the usual heteroskedasticity consistent variance matrix. For the remaining groups, G2 and G4 respectively indicate partitioning the data into two and four geographic regions.[23] T3, T6, and T32 divide the time series into three 128-month periods, six 64-month periods, or thirty-two 12-month periods. "G4 x T3" then indicates a group structure where observations in region one in time period one belong to the same group, observations in region two in time period one belong to the same group, etc. For all simulations, we use the full sample with 49 states and 384 time periods noting that this leaves 383 time series observations for use in estimation with first-differences and produces many grouping schemes with different though similar number of observations per groups. All results are based on 1000 simulation replications.

The simulations show that tests based on standard errors which ignore any of the sources of correlation (IID, Heteroskedasticity, clustering with either state or month as a grouping variable)

[22]We scaled x so that the standard deviation of x matches the empirical standard deviation of $\log(y_{st})$.

[23]For G4, we use the four census regions; Northeast, Midwest, South, and West; but modify them slightly by taking Delaware, Maryland, and Washington D.C. from the south and adding them to the Northeast. For G2, we essentially split the country into East and West at the Mississippi river but include Wisconsin and Illinois in the West.

perform poorly across the designs considered with the exception of clustering by month in the first-difference AR(1) since the treatment has very little intertemporal correlation after differencing. We also see that grouping strategies that use a large number of groups fare quite poorly in these designs, again with the exception of the first-differenced AR(1) results. On the other hand, the conservative grouping strategies; T3, G2, and G4; appear to perform well across all simulation designs. G2 x T3 also does quite well across all simulation designs reported though it suffers from larger size distortions than the more conservative strategies.

In practice, one might prefer tests with moderate size distortions if they are sufficiently more powerful than tests with size closer to the nominal level. We note that power should increase with degrees of freedom of the fixed-G asymptotic t distribution as increases in degrees of freedom decrease the appropriate critical values. Since relevant critical values of a t-distribution are highly concave in the degrees of freedom, there will be rapidly diminishing returns to increasing the degrees of freedom. Figure 2 plots power curves for G2, T3, and G4. In the figure, the solid curve plots power for G4, the crossed line plots power for G2, and the line with circles plots power for T3. The figure clearly illustrates the power gain from moving to configurations with more groups. Moving from G2 to T3 to G4, sizes are similar, but the power from G4 is substantially higher than that of T3, and G2, which uses Cauchy critical values, has very low power relative to both T3 and G4.

These simulation results illustrate the potential for inference procedures that fail to account for both spatial and inter-temporal correlation in panel data to produce extremely misleading results. Probably the most common current inference approaches in panel data are based on using standard errors clustered at the cross-sectional unit of observation, state in our simulation example, which allows general inter-temporal correlation but essentially ignores cross-sectional correlation. Our simulations based on actual unemployment data suggest that this has the potential to produce substantial size distortions in tests of hypotheses. Another popular approach is to treat observations as if they belong to the same group if they are from the same cross-sectional unit or the same time series unit, which corresponds to our "state/month" results. The simulation results also

suggest that inference based on this group structure may have substantial size distortions in the presence of inter-temporal and cross-sectional correlation. While we have not dealt with optimal group selection, the results suggest that one needs to be very conservative when defining groups to produce inference statements that have approximately correct coverage or size. The fact that in all cases we find that one should use a quite a small number of groups to produce inference with size close to the nominal level suggests that one might wish to consider estimation methods that more efficiently use the available information and that there may be gains to more carefully considering group construction. We leave exploring these issues to future research.

4.3. **2SLS Simulations and Comparison to Ibragimov and Müller (2006)**

This section presents a set of simulation experiments contrasting performance of t-tests using our approach, referred to below as BCH, with that of IM. Specifically, the IM t-statistic is formed by constructing a different point estimate of a parameter θ within each group using only observations within that group, call it $\hat{\theta}_g$. Letting $\bar{\theta}_G$ denote the cross group average point estimator, $\bar{\theta}_G = \frac{1}{G}\sum_g \hat{\theta}_g$, the IM test statistic for the null hypothesis that $\theta = \theta_0$ is then simply

$$t_{IM} = \frac{\bar{\theta}_G - \theta_0}{\frac{1}{\sqrt{G}}\sqrt{\frac{1}{G-1}\sum_g (\hat{\theta}_g - \bar{\theta}_G)^2}}$$

.

As we note in Section 3.1, we conjecture that the IM approach will perform better in data that feature pronounced heterogeneity in regressor variances, while our approach should perform better when the estimator being used exhibits appreciable finite sample bias. Since both papers advocate the use of a small number of groups to obtain correctly sized tests, we will examine their performance using data sets that have $T = 100$ observations and $G = 4$ groups. Our simulation exercises use three basic DGPs, the first of which is the following regression model with Gaussian regressors and error terms following independent AR(1) processes with $\rho = \frac{1}{2}$,

$$
\begin{aligned}
Y_t &= b_0 + b_1 X_t + \varepsilon_t \\
\varepsilon_t &= \rho \varepsilon_{t-1} + u_t \\
X_t &= \rho X_{t-1} + v_t
\end{aligned}
$$

where v_t is IID $N(0,1)$ and independent of u_t which is also standard normal. The true parameter values are $b_0 = 0$ and $b_1 = 1$. Initial conditions are drawn from stationary distributions and point estimates obtained by OLS.

In our next experiments, we consider inference using the 2SLS estimator which allows us to highlight the impact of finite sample bias on the procedures' performance. Specifically, we employ the following simple design:

$$
\begin{aligned}
Y_t &= b_0 + b_1 X_t + \varepsilon_t \\
\varepsilon_t &= \rho \varepsilon_{t-1} + u_t \\
X_t &= \Pi[11...1]' Z_t + v_t \\
&\quad \begin{bmatrix} u_t \\ v_t \end{bmatrix} \sim \text{ IID } N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \frac{1}{1-\rho^2} \begin{bmatrix} 1 & -.8 \\ -.8 & 1 \end{bmatrix} \right) \\
Z_t &= \rho Z_{t-1} + W_t, \quad W_t \sim N\left(0, \frac{1}{k} I_k\right), \quad \text{where} \dim(Z) = k.
\end{aligned}
$$

We keep ρ fixed at $\frac{1}{2}$ and study the impact of varying instrument strength ($\Pi = \frac{1}{2}, 1, 2$) and number of instruments ($k = 3, 6$), which leads to varying degrees of bias in the numerators of test statistics.

In our final design, we also consider the impact of pronounced heterogeneity in subsample (group) variances. We consider a DGP where the first element of the instrument vector, Z_{1t}, is strictly stationary but features rare jumps. Hence in samples of size $T = 100$, Z_{1t} will exhibit

pronounced disparity in within-group sample variances as a function of a large jump size parameter, J. We leave the elements of Z_t beyond the first unchanged from the specification above. Our process for Z_{1t} is

$$Z_{1t} = 1(V_t = 0)S_t + V_t$$

$$V_t = \begin{cases} 0 \text{ with prob } 96\% \\ +J \text{ with prob } 2\% \\ -J \text{ with prob } 2\% \end{cases}$$

$$S_t = \begin{cases} \rho S_{t-1} + w_t \text{ when } V_{t-1} = 0 \\ \rho S_{t-2} + w_t \text{ when } V_{t-1} = \pm J \end{cases}$$

This process intermittently has fluctuations due to the V_t component and then resets itself the period after the $\pm J$ shock.

Table 3 presents our simulation results for the linear model as well as the homogeneous 2SLS design with 3 or 6 instruments and varying instrument strength with Π being 2, 1, and 1/2. The columns headed Bias and RMSE report the simulation bias and root mean squared error for the numerators of the associated t-statistics. The column heading Size refers to simulation rejection frequency 5% level t-tests under a correct null hypothesis. The first panel presents our results for the regression model with serial correlation but no endogeneity: both methods have numerators with small bias and perform well in terms of size. Results for 2SLS t-tests are, in contrast, very dependent on the amount of bias possessed by the 2SLS point estimators, which increases as Π decreases or k increases. Our benchmark specification ($\Pi = 2$, $k = 3$) was deliberately chosen to generate comparable sizes for BCH and IM. As we move away from the benchmark by decreasing instrument strength or increasing the number of instruments, bias increases and IM begins to suffer large size distortions while BCH remains much closer to having correct size.

Table 4 presents a specification designed to investigate the importance of group variance heterogeneity and instrument strength upon test performance. Our 'intermittent jump' process is

used with differing values of J =5, 10, and 20. These values of J result in variation in the sample variance of Z_{1t} across groups with large variation for J = 10 and 20. For the strong instrument setting, $\Pi = 2$, shown in the top panel of the table, BCH suffers from size distortions that increase as the heterogeneity in sample variances grows. As anticipated, this distortion is not present for IM. However, as the final two panels of the table illustrate, when instrument strength declines, the increases in numerator biases again lead to size distortions in the IM t-test, so that neither estimator is superior in all cases. Hence, as stated above, we believe these are complementary approaches.

5. **Conclusion**

In this paper, we use the clustered covariance matrix estimator (CCE) to perform inference about regression parameters when data are weakly dependent. We allow for general forms of dependence that cover time series, spatial, and panel data. We show that inference based on the CCE is valid in these contexts despite the fact that data do not follow a grouped structure under weak dependence.

We establish our results using an asymptotic sequence in which the number of groups is fixed and the number of observations per group goes to infinity. Under this sequence, the CCE is not consistent but converges in distribution to a nondegenerate random variable. In this case, standard t and Wald tests based on the CCE converge in distribution to ratios of random variables that reflect the estimation uncertainty for the covariance matrix. This result is similar to that obtained in Kiefer and Vogelsang (2002, 2005) (KV) who consider inference using a usual HAC estimator in a time series context. Under mild homogeneity conditions, we show that the limiting distributions of our t and Wald statistics are proportional to standard t and F distributions, which results in extremely simple-to-implement testing procedures. Simulation results show that our asymptotic approximations perform quite well relative to using HAC with the usual asymptotic approximation and are on par with results obtained using the KV approximation though we lose some power to

KV in some designs. In a recent paper, Sun, Phillips, and Jin (2008) have shown that the KV "fixed-b" approach provides an asymptotic refinement relative to the usual asymptotic approach for time series HAC in a Gaussian location model. We conjecture that our results also provide such a refinement.

The simulations show that tests and confidence intervals based on the CCE and the fixed-G approximations have size and coverage close to the nominal level under sensible designs with intertemporal correlation, spatial correlation, and a panel with a combination of the two. In all of our simulation results, correctly-sized tests are only produced when one uses a relatively small number of groups when there is non-negligible correlation in the data. The desirability of a small number of groups further demonstrates the usefulness of the fixed-G results. Finally, it bears repeating that inference based on the CCE is extremely tractable computationally and that the fixed-G reference distributions are standard, making implementing the procedure straightforward in practice.

An important unanswered question is smoothing parameter selection, which corresponds to choice of groups in our context. In principle, we could consider smoothing parameter selection for the CCE based on minimizing mean squared error (MSE) for estimating the asymptotic variance; see, e.g. Andrews (1991). However, in much of applied research, the chief reason that one wishes to estimate a covariance matrix is in order to perform inference about estimated model parameters. Minimizing MSE of the covariance matrix estimator will not necessarily translate to good inference properties. Our simulation results suggest that one needs to use quite a large smoothing parameter (resulting in a covariance estimate with small degrees of freedom) to control the size of a test when using a HAC or CCE. It appears that having an estimator with smaller bias than would be MSE optimal for estimating the covariance matrix itself is important for tests to have approximately correct size. This is consistent with Sun, Phillips, and Jin (2008), who consider this problem in the context of Gaussian location model in a time series and show that the rate of increase for the optimal smoothing parameter chosen by trading off size and power is much faster than the rate for

minimizing MSE of the variance estimator. An interesting direction for future research would be to adapt the arguments of Sun, Phillips, and Jin (2008) to the present context.

6. Appendix A. Proofs of Propositions

Throughout the appendix, we suppress the dependence of smoothing parameters and estimators on N, writing, for example, $\widehat{V}_N$ as $\widehat{V}$ and the number of groups and the number of elements per group simply as G and L. We use CMT to denote the continuous mapping theorem and CS to denote the Cauchy-Schwarz inequality. We use C as a generic constant whose value may change depending on the context.

6.1. Proof of Proposition 1

The proof of the proposition is based on the following expression for $\hat{V}$:

$$\begin{aligned}\hat{V} = \frac{1}{G}\sum_{g=1}^{G}\Bigg\{ & \frac{x_g'\varepsilon_g}{\sqrt{L}}\frac{\varepsilon_g' x_g}{\sqrt{L}} \\ & - \frac{x_g' x_g}{L}\left(\sum_{h=1}^{G}\frac{x_h' x_h}{L}\right)^{-1}\left(\sum_{h=1}^{G}\frac{x_h'\varepsilon_h}{\sqrt{L}}\right)\frac{\varepsilon_g' x_g}{\sqrt{L}} - \frac{x_g'\varepsilon_g}{\sqrt{L}}\left(\sum_{h=1}^{G}\frac{x_h'\varepsilon_h}{\sqrt{L}}\right)'\left(\sum_{h=1}^{G}\frac{x_h' x_h}{L}\right)^{-1}\frac{x_g' x_g}{L} \\ & + \frac{x_g' x_g}{L}\left(\sum_{h=1}^{G}\frac{x_h' x_h}{L}\right)^{-1}\left(\sum_{h=1}^{G}\frac{x_h'\varepsilon_h}{\sqrt{L}}\right)\left(\sum_{h=1}^{G}\frac{x_h'\varepsilon_h}{\sqrt{L}}\right)'\left(\sum_{h=1}^{G}\frac{x_h' x_h}{L}\right)^{-1}\frac{x_g' x_g}{L}\Bigg\}.\end{aligned}$$

Let $B_g \sim N(0, I_k)$ denote a random k-vector and $\Omega_g = \Lambda_g\Lambda_g'$. Define matrices $\mathbf{Q}$ and $\mathbf{S}$ such that $\mathbf{Q} = \sum_g Q_g$ and $\mathbf{S} = \sum_g \Lambda_g B_g$. Note that Assumption 3 implies $\mathbf{Q} = GQ$ while Assumption 4 implies $\Lambda_g = \Lambda$, and therefore $\mathbf{S} = \Lambda\sum_g B_g$. The following three random variables will be limits of $\hat{V}$ under Assumptions 1-2, 1-3, and 1-4 respectively:

$$V_A = \frac{1}{G}\sum_g\left[\Lambda_g B_g B_g'\Lambda_g' - Q_g\mathbf{Q}^{-1}\mathbf{S}B_g'\Lambda_g' - \Lambda_g B_g\mathbf{S}'\mathbf{Q}^{-1}Q_g + Q_g\mathbf{Q}^{-1}\mathbf{S}\mathbf{S}'\mathbf{Q}^{-1}Q_g\right]$$

$$V_B = \frac{1}{G}\sum_g\left[\Lambda_g B_g B_g'\Lambda_g' - \frac{1}{G}\mathbf{S}B_g'\Lambda_g' - \frac{1}{G}\Lambda_g B_g\mathbf{S}' + \frac{1}{G^2}\mathbf{S}\mathbf{S}'\right]$$

$$V_C = \frac{1}{G}\Lambda\left[\sum_g B_g B_g' - \frac{1}{G}\left(\sum_g B_g\right)\left(\sum_g B_g'\right)\right]\Lambda'.$$

Note that V_B is equivalent to V_A under $Q_g = Q$, and that V_C is equivalent to V_B under $\Lambda_g = \Lambda$.

(i) $\hat{V} \xrightarrow{d} V_A$ is immediate from Lemma 1 and the CMT. It is also immediate from Lemma 1 and the CMT that $\sqrt{L}(\hat{\beta} - \beta) \xrightarrow{d} \mathbf{Q}^{-1}\mathbf{S}$. The result is then obvious from the CMT.

(ii) $\hat{V} \xrightarrow{d} V_A$ is again immediate under Lemma 1 and the CMT, and $V_A = V_B$ is immediate under Assumption 3 plugging in $\mathbf{Q} = GQ$. Under Assumptions 1-3 and H_0, we have that

$$\sqrt{N}\left(R\widehat{\beta} - r\right) \xrightarrow{d} \sqrt{G}RQ^{-1}\frac{1}{G}\sum_g \Lambda_g B_g$$
$$R\hat{Q}^{-1}\hat{V}\hat{Q}^{-1}R' \xrightarrow{d} RQ^{-1}V_B Q^{-1}R'$$

We can write the RHS of the second line as

$$RQ^{-1}\left(\frac{1}{G}\sum_g\left[\Lambda_g B_g B_g'\Lambda_g - \frac{1}{G}\mathbf{S}B_g'\Lambda_g' - \frac{1}{G}B_g\Lambda_g\mathbf{S} + \frac{1}{G^2}\mathbf{S}\mathbf{S}'\right]\right)Q^{-1}R'$$
$$= \frac{1}{G}\sum_g\left[RQ^{-1}\Lambda_g B_g B_g'\Lambda_g'^{-1}Q^{-1}R' - \frac{1}{G}\widetilde{\mathbf{S}}B_g'\Lambda_g'^{-1}Q^{-1}R' - \frac{1}{G}RQ^{-1}B_g\Lambda_g\widetilde{\mathbf{S}} + \frac{1}{G^2}\widetilde{\mathbf{S}}\widetilde{\mathbf{S}}'\right],$$

where $\widetilde{\mathbf{S}} = \sum_g RQ^{-1}\Lambda_g B_g$. Letting $B_{1,g} \sim N(0,1)$ and supposing R is $1 \times k$, we therefore have

$$\hat{t} \xrightarrow{d} \sqrt{G}\frac{\frac{1}{G}\sum_g \lambda_g B_{1,g}}{\sqrt{\frac{1}{G}\sum_g\left[\lambda_g B_{1,g} - \left(\frac{1}{G}\sum_g \lambda_g B_{1,g}\right)\right]^2}}$$
$$= \sqrt{\frac{G}{G-1}}\left(\sqrt{G}\frac{\frac{1}{G}\sum_g \lambda_g B_{1,g}}{\sqrt{\frac{1}{G-1}\sum_g\left[\lambda_g B_{1,g} - \left(\frac{1}{G}\sum_g \lambda_g B_{1,g}\right)\right]^2}}\right),$$

where $\lambda_g^2 = RQ^{-1}\Lambda_g\Lambda_g'Q^{-1}R'$. The result then follows immediately from Theorem 1 of Ibragimov and Müller (2006); see also Bakirov and Székely (2005).

(iii) $\hat{V} \xrightarrow{d} V_A$ is again immediate under Lemma 1 and the CMT, and $V_A = V_C$ is immediate under Assumptions 3 and 4 plugging in $\mathbf{Q} = GQ$ and $\Lambda_g = \Lambda$. Under Assumptions 1-4 and under

H_0, we also immediately have

$$\sqrt{N}\left(R\widehat{\beta}-r\right) \xrightarrow{d} \sqrt{G}RQ^{-1}\Lambda\left(\frac{1}{G}\sum_g B_g\right)$$
$$R\hat{Q}^{-1}\hat{V}\hat{Q}^{-1}R' \xrightarrow{d} RQ^{-1}V_C Q^{-1}R'$$

Let R be $1\times k$ and r be a scalar. In this case, $\lambda^2 = RQ^{-1}\Lambda\Lambda' Q^{-1}R'$ is a scalar, and letting $B_{1,g}$ be a scalar standard normal r.v., we have

$$\hat{t} \xrightarrow{d} \frac{\lambda G^{-1/2}\sum_g B_{1,g}}{\sqrt{\lambda^2 G^{-1}\left[\sum_g B_{1,g}^2 - \frac{1}{G}\left(\sum_g B_{1,g}\right)^2\right]}} = \sqrt{\frac{G}{G-1}}\frac{B_{1,G}}{\sqrt{(\sum_g B_{1,g}^2 - B_{1,G}^2)/(G-1)}},$$

where $B_{1,G} \equiv G^{-1/2}\sum_g B_{1,g} \sim N(0,1)$ and $\sum_g B_{1,g}^2 - B_{1,G}^2 \sim \chi^2_{G-1}$ are independent.

It follows that $\hat{t} \xrightarrow{d} \sqrt{\frac{G}{G-1}}t_{G-1}$. The result for $\hat{F}$ is similar using Rao (2002) Chapter 8b. ■

6.2. **Proof of Lemma 1 with Spatial Dependence**

We provide a proof for the m-dimensional case.

Assumption 1.(iii)-(iv) immediately imply $\frac{1}{L}x_g'x_g \xrightarrow{p} Q_g$ which follows from Jenish and Prucha (2007) Theorem 3 for all $g = 1, ..., G$ from which Lemma 1.(i) follows. Next, Assumptions 1.(iii)-(iv) imply the conditions of Jenish and Prucha (2007) Theorem 1 for $\frac{1}{\sqrt{L}}x_g'\varepsilon_g$ for $g = 1, ..., G$ from which it follows that the array $\left(\frac{1}{\sqrt{L}}x_1'\varepsilon_1, ..., \frac{1}{\sqrt{L}}x_G'\varepsilon_G\right)' \xrightarrow{d} Z = N(0, W)$ where Z follows a multivariate normal distribution with variance matrix W. It now remains to be shown that W is block diagonal when grouped with blocks corresponding to covariances across groups.

Let generic groups be denoted g and h. An off-diagonal block of W corresponds to the limit as $L \to \infty$ of

$$\frac{1}{L}E\left[\left(\sum_{s\in g} x_s\varepsilon_s\right)\left(\sum_{r\in h} x_r\varepsilon_r\right)\right] \leq \frac{1}{L}\sum_{s\in g}\sum_{r\in h}|Ex_s\varepsilon_s x_r\varepsilon_r|.$$

which needs to be shown to go to 0. Let us call the object that needs to be shown to vanish R_L:

$$R_L = \frac{1}{L}\sum_{s\in g}\sum_{r\in h}|Ex_s\varepsilon_s x_r\varepsilon_r|$$

We first note that the largest number of d^{th} order neighbors for any set of k points is $C(m)kd$ where $C(m)$ is a constant that depends on the dimension of the index set. Under the boundary condition in Assumption 2.(iv), there are at most $CL^{(m-1)/m}$ observations on the boundary of any set g. In addition, the boundary points are contiguous under Assumption 2.(iii). In counting the number of neighbors, it is useful to think of each group as a collection of 'contour sets.' First, the boundary, then the set of interior points that are one unit from the boundary, then the interior points two units from the boundary and so on. For d^{th} order neighbors, there are d different pairs of contour sets that the neighbors can reside in. For example, a pair of second-order neighbors must contain one point on the boundary of either g or h and another point in the first contour off the boundary of the other set. In addition, the largest any contour set can be is the maximum size of the boundary. This allows us to bound the maximum number of pairs with any given contour set memberships by the maximum number of first-order neighbors, $C(m)L^{(m-1)/m}$. Combining these two observations, we can bound the maximum number of d^{th} order neighbors by $C(m)L^{(m-1)/m}d$.

Using this bound, we can write

$$R_L = \frac{1}{L}\sum_{s\in g}\sum_{r\in h}|Ex_s\varepsilon_s x_r\varepsilon_r| \leq \frac{1}{L}\Delta C(m)L^{(m-1)/m}\sum_{d=1}^{\infty} d\alpha_{1,1}(d)^{\frac{\delta}{2+\delta}} = O(L^{-1/m})$$

using the moment conditions in Assumption 1.(iii) and a standard mixing inequality, e.g. Jenish and Prucha (2007) or Bolthausen (1982) Lemma 1, to obtain the inequality and the mixing rate assumptions in Assumption 1.(iii) to show that $\sum_{d=1}^{\infty} d\alpha_{1,1}(d)^{\frac{\delta}{2+\delta}}$ converges. It follows immediately that Assumptions 1 and 2 imply the conclusion of Lemma 1. ■

7. Appendix B. Unequal Group Sizes

In this section, we present results for the case where group sizes are not asymptotically equivalent. Results are presented without proof but follow from the same arguments as Proposition 1.

To establish the results, we replace Assumption 2.(ii) with a simple condition that allows for unequal group sizes such that no groups are dominant and under which Lemma 1 will continue to be satisfied:

Assumption 5. *For all g, $|\mathcal{G}_g| = L_{g,N}$, and $L_{g,N}/L_N \to \rho_g$ where $L_N = \frac{1}{G}\sum_g L_{g,N}$.*

We note that no modification of Lemma 1, outside of replacing Assumption 2.(ii) with Assumption 7, is necessary. We also consider cases where we add homogeneity to the model by replacing Assumptions 3 and 4 with Assumptions 8 and 9.

Assumption 6. *For all g, $Q_g \equiv \rho_g Q$.*

Assumption 7. *For all g, $\Omega_g \equiv \rho_g \Omega$.*

Once again, Assumptions 8 and 9 are implied by covariance stationarity of the individual observations but may also be satisfied even if covariance stationarity is violated.

With the assumptions modified, we state the analog of Proposition 1 for unequal group sizes. In case 3, we only state the result for the Wald statistic, $\widehat{F}$, to conserve space.

Proposition 2. *Suppose $\{\mathcal{G}_g\}$ is defined such that $L_N \to \infty$ and G is fixed as $N \to \infty$ and that Lemma 1 holds. Let $B_g \sim N(0, I_k)$ denote a random k-vector and $\Omega_g = \Lambda_g \Lambda_g'$. Define matrices $\mathbf{Q}$ and $\mathbf{S}$ such that $\mathbf{Q} = \sum_g Q_g$ and $\mathbf{S} = \sum_g \Lambda_g B_g$. Then,*

i. $\hat{V}_N \xrightarrow{d} V_A = \frac{1}{G}\sum_g \left[\Lambda_g B_g B_g' \Lambda_g' - Q_g \mathbf{Q}^{-1}\mathbf{S} B_g' \Lambda_g' - \Lambda_g B_g \mathbf{S}'\mathbf{Q}^{-1} Q_g + Q_g \mathbf{Q}^{-1}\mathbf{S}\mathbf{S}'\mathbf{Q}^{-1} Q_g\right]$, *and under H_0,*

$$\hat{t} \xrightarrow{d} \frac{\sqrt{G} R\mathbf{Q}^{-1}\mathbf{S}}{\sqrt{R(\mathbf{Q}/G)^{-1} V_A (\mathbf{Q}/G)^{-1} R'}} \qquad \text{and}$$

$$\hat{F} \xrightarrow{d} G\mathbf{S}'\mathbf{Q}^{-1} R' \left[R(\mathbf{Q}/G)^{-1} V_A (\mathbf{Q}/G)^{-1} R'\right]^{-1} R\mathbf{Q}^{-1}\mathbf{S}.$$

ii. *Define* $\hat{\beta}_w$, $\hat{Q}_w$, *and* $\hat{V}_w$ *as the respective weighted least squares estimates where the observations in group* g *are weighted by* $\sqrt{L_N/L_{g,N}}$. *Define* $\hat{t}_w$ *using the WLS estimates in place of* $\hat{\beta}_N$, $\hat{Q}$, *and* $\hat{V}_N$. *If Assumption 8 is also satisfied,* $\hat{t}_w \xrightarrow{d} \sqrt{\frac{G}{G-1}} t^*_{G-1}$ *under* H_0 *where* t^*_{G-1} *satisfies*

$$P\left(|t^*_{G-1}| > c_{G-1}(\alpha)\right) \leq \alpha$$

for $c_{G-1}(\alpha)$ *the usual critical value for an* $\alpha-$*level two-sided t-test based on a t-distribution with* $G-1$ *degrees of freedom for any* $\alpha \leq 2\Phi(-\sqrt{3})$ *and for any* $\alpha \leq 0.1$ *if* $2 \leq G \leq 14$.

iii. *Let* $B_g \sim N(0, I_q)$ *be independent across* g *where* R *is* $q \times k$; *and define* $S_1 = \sum_g \rho_g^{1/2} B_g$ *and* $S_2 = \sum_g \rho_g^{3/2} B_g$. *If Assumptions 8 and 9 are satisfied,* $\hat{F} \xrightarrow{d} S_1'[\sum_g \rho_g B_g B_g' - S_2 S_1' - S_1 S_2' + \sum_g \rho_g^2 S_1 S_1']^{-1} S_1$ *under* H_0.

Proposition 3.i is identical to Proposition 1.i which already allowed for heterogeneity. To establish the analog of Proposition 1.ii, we use a weighted estimator that reweights so the design matrix heterogeneity cancels out. This reweighting places additional weight on groups with small numbers of observations and smaller weight on groups with larger numbers of observations which may seem undesirable. We note that this is essentially what is done in IM as well since that use unweighted estimators formed within groups; that is the estimator from a small group receives exactly as much weight as an estimate from a large group in the IM scheme. It is interesting that this same apparently undesirable reweighting produces additional robustness to heterogeneity in our approach as well as allowing IM to establish their results. Finally, the limiting distribution under homogeneity of both the design matrices and the scores but with unequal group sizes is non-standard and depends on the group sizes. Since these are readily obtained and the distribution is otherwise free of nuisance parameters, it can easily be simulated and used to conduct inference in cases in which one is uncomfortable forming approximately equal-sized groups.

8. **Appendix C. Nonlinear Models**

We provide a sketch of the modifications of our regularity conditions for the fixed-G result to hold for m-estimators.

Suppose that

$$\widehat{\theta} = \arg\max_{\theta} \frac{1}{N} \sum_i f(z_{s_i}; \theta)$$

where $\lim_{N\to\infty} \frac{1}{N} \sum_i E[f(z;\theta)]$ is maximized at some parameter value θ_0. For simplicity, assume also that $f(z;\theta)$ is twice-continuously differentiable in θ. We will have that $\widehat{\theta} \xrightarrow{p} \theta_0$ and $\sqrt{N}(\widehat{\theta} - \theta_0) \xrightarrow{d} \Gamma^{-1} N(0, V)$ where $V = \lim_{N\to\infty} \text{Var}\left[\frac{1}{\sqrt{N}} \sum_i \frac{\partial}{\partial\theta} f(z_{s_i}; \theta_0)\right]$ and $\Gamma = \lim_{N\to\infty} \text{E}\left[\frac{1}{N} \sum_i \frac{\partial^2}{\partial\theta\partial\theta'} f(z_{s_i}; \theta_0)\right]$ under standard regularity conditions; see, for example, Wooldridge (1994) in the time series case and Jenish and Prucha (2007) in the spatial case.[24]

Let $D(z_{s_i};\theta) = \frac{\partial}{\partial\theta} f(z_{s_i};\theta)$ be a $k \times 1$ vector and let $D_g(\theta) = \sum_{i\in\mathcal{G}_g} D(z_{s_i};\theta)$ be the $k \times 1$ vector defined by summing the first derivatives within group g for $g = 1, ..., G$. Also, define $\Gamma_g(\theta) = \sum_{i\in\mathcal{G}_g} \frac{\partial^2}{\partial\theta\partial\theta'} f(z_{s_i};\theta)$. Then the clustered estimator of V would be given by

$$\widehat{V} = \frac{1}{N} \sum_{g=1}^{G} D_g(\widehat{\theta}) D_g(\widehat{\theta})'.$$

We can then follow the usual procedure in the HAC literature and linearize $Df_g(\widehat{\theta})$ around the true parameter θ_0. This gives

$$\widehat{V}_N = \frac{1}{N} \sum_{g=1}^{G} \Big[D_g(\theta_0) D_g(\theta_0)' + \Gamma_g(\bar{\theta})(\widehat{\theta} - \theta_0) D_g(\theta_0)' + D_g(\theta_0)(\widehat{\theta} - \theta_0)' \Gamma_g(\bar{\theta}) + \Gamma_g(\bar{\theta})(\widehat{\theta} - \theta_0)(\widehat{\theta} - \theta_0)' \Gamma_g(\bar{\theta}) \Big]$$

[24] Jenish and Prucha (2007) provides conditions for uniform laws of large numbers and central limit theorems. To show consistency and asymptotic normality, these results would need to be combined with standard consistency and asymptotic normality results for m-estimators as in Newey and McFadden (1994).

where $\bar{\theta}$ is an intermediate value. By standard arguments, we can also write that

$$\widehat{\theta} - \theta_0 = -\left[\sum_{g=1}^{G} \Gamma_g(\bar{\theta})\right]^{-1} \sum_g D_g(\theta_0)$$

with $\bar{\theta}$ an intermediate value. Substituting this expression into $\widehat{V}_N$, we have

$$\begin{aligned}
\widehat{V}_N = \frac{1}{N}\sum_{g=1}^{G} \Bigg[& D_g(\theta_0) D_g(\theta_0)' \\
& - \Gamma_g(\bar{\theta}) \left(\left[\sum_{r=1}^{G} \Gamma_r(\bar{\theta})\right]^{-1} \sum_{r=1}^{G} D_r(\theta_0)\right) D_g(\theta_0)' \\
& - D_g(\theta_0) \left(\left[\sum_{r=1}^{G} \Gamma_r(\bar{\theta})\right]^{-1} \sum_{r=1}^{G} D_r(\theta_0)\right)' \Gamma_g(\bar{\theta}) \\
& + \Gamma_g(\bar{\theta}) \left(\left[\sum_{r=1}^{G} \Gamma_r(\bar{\theta})\right]^{-1} \sum_{r=1}^{G} D_r(\theta_0)\right) \left(\left[\sum_{r=1}^{G} \Gamma_r(\bar{\theta})\right]^{-1} \sum_{r=1}^{G} D_r(\theta_0)\right)' \Gamma_g(\bar{\theta}) \Bigg].
\end{aligned}$$

Looking at this expression, we see that $D_g(\theta_0)$ is playing the same role as $x_g'\varepsilon_g$ in Section 3.1 and $\Gamma_g(\bar{\theta})$ is playing the same role as $x_g'x_g$. It will follow immediately that the appropriate sufficient condition analogous to Lemma 1 above will have that

$$\frac{1}{\sqrt{L_N}} \left(D_1(\theta_0), ..., D_{G_N}(\theta_0)\right)' \xrightarrow{d} N(0, W)$$

where W is block diagonal with off-diagonal blocks equal to matrices of zeros and diagonal blocks equal to Ω_g where $\Omega_g = \lim_{L_N \to \infty} \text{Var}\left[\frac{1}{\sqrt{L_N}} D_g(\theta_0)\right]$ and that $\sup_{\theta \in \Theta} \| \frac{1}{L_N} \Gamma_g(\theta) - \Gamma_g^*(\theta) \| \xrightarrow{p} 0$ where $\Gamma_g^*(\theta_0)$ is nonsingular for all $g = 1, ..., G_N$. Primitive conditions for the first condition can be found in any standard reference for central limit theorems; see, for example, Jenish and Prucha (2007) for spatial processes and White (2001) for time series processes.[25] The second condition is a uniform convergence condition for the Hessian matrix

[25] Additional conditions regarding the group structure such as those in Assumption 2 would also have to be added to verify the block diagonality. This could be demonstrated as in Appendix 6.2.

for which a variety of primitive conditions can be found, e.g. Jenish and Prucha (2007) or Wooldridge (1994).

References

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica 59*(3), 817–858.

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics 49*(4), 431–434.

Bakirov, N. K. and G. J. Székely (2005). Student's t-test for gaussian scale mixtures. *Zapinski Nauchnyh Seminarov POMI 328*, 5–19.

Bartlett, M. S. (1950). Periodogram analysis and continuous spectra. *Biometrika 37*, 1–16.

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

Bester, C. A., T. G. Conley, and C. B. Hansen (2010). Inference for dependent data using cluster covariance estimators. Available at SSRN: http://ssrn.com/abstract=1708263.

Bester, C. A., T. G. Conley, C. B. Hansen, and T. J. Vogelsang (2008). Fixed-b asymptotics for spatially dependent robust nonparametric covariance matrix estimators. Mimeo.

Bolthausen, E. (1982). On the central limit theorem for stationary mixing random fields. *The Annals of Probability 10*, 1047–1050.

Conley, T. G. (1996). *Econometric Modelling of Cross-Sectional Dependence*. Ph.D. Dissertation, University of Chicago.

Conley, T. G. (1999). Gmm estimation with cross sectional dependence. *Journal of Econometrics 92*, 1–45.

Conley, T. G. and E. A. Ligon (2002). Economic distance, spillovers, and cross country comparisons. *Journal of Economic Growth 7*, 157–187.

Fama, E. F. and J. MacBeth (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy 81*, 607–636.

Foote, C. L. (2007). Space and time in macroeconomic panel data: Young workers and state-level unemployment revisited. Federal Reserve Bank of Boston Working Paper No. 07-10.

Hansen, C. B. (2007). Asymptotic properties of a robust variance matrix estimator for panel data when t is large. *Journal of Econometrics 141*, 597–620.

Ibragimov, R. and U. K. Müller (2006). t-statistic based correlation and heterogeneity robust inference. Mimeo.

Jansson, M. (2004). The error in rejection probability of simple autocorrelation robust tests. *Econometrica 72*(3), 937–946.

Jenish, N. and I. Prucha (2007). Central limit theorems and uniform laws of large numbers for arrays of random fields. Mimeo.

Kelejian, H. H. and I. Prucha (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review 40*, 509–533.

Kelejian, H. H. and I. Prucha (2001). On the asymptotic distribution of the moran i test statistic with applications. *Journal of Econometrics 104*, 219–257.

Kiefer, N. M. and T. J. Vogelsang (2002). Heteroskedasticity-autocorrelation robust testing using bandwidth equal to sample size. *Econometric Theory 18*, 1350–1366.

Kiefer, N. M. and T. J. Vogelsang (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory 21*, 1130–1164.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. New York: Springer-Verlag.

Lee, L.-f. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial econometric models. *Econometrica 72*, 1899–1926.

Lee, L.-f. (2007a). Gmm and 2sls estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics 137*, 489–514.

Lee, L.-f. (2007b). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics 140*, 333–374.

Liang, K.-Y. and S. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In R. F. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics. Volume 4*. Elsevier: North-Holland.

Newey, W. K. and K. D. West (1987). A simple, positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*(3), 703–708.

Pulvino, T. C. (1998). Do asset fire sales exist? an empirical investigation of commercial aircraft transactions. *Journal of Finance 54*(3), 929–978.

Rao, C. R. (2002). *Linear Statistical Inference and Its Application*. Wiley-Interscience.

Shimer, R. (2001). The impact of young workers on the aggregate labor market. *Quarterly Journal of Economics 116*, 969–1008.

Stata Corporation (2007). *Stata User's Guide Release 10*. College Station, Texas: Stata Press.

Stock, J. H. and M. W. Watson (2008). Heteroskedasticity-robust standard errors for fixed effects panel data regression. *Econometrica 76*(1), 155–174.

Sun, Y., P. C. B. Phillips, and S. Jin (2008). Optimal bandwidth selection in heteroskedasticity-autocorrelation robust testing. *Econometrica 76*(1), 175–194.

White, H. (2001). *Asymptotic Theory for Econometricians* (Revised ed.). San Diego: Academic Press.

Wooldridge, J. M. (1994). Estimation and inference for dependent processes. In R. F. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics. Volume 4*. Elsevier: North-Holland.

Wooldridge, J. M. (2003). Cluster-sample methods in applied econometrics. *American Economic Review 93*(2), 133–188.

Table 1. Simulation Results. T-test Rejection Rates for 5% Level Tests

		Time Series			Spatial		
	Ref. Dist.	ρ=0.0	ρ=0.5	ρ=0.8	γ=0.0	γ=0.3	γ=0.6
IID	N(0,1)	0.050	0.126	0.340	0.054	0.388	0.556
Heteroskedasticity	N(0,1)	0.057	0.138	0.364	0.054	0.386	0.560
Bartlett-Large H	N(0,1)	0.179	0.203	0.258	0.692	0.708	0.712
Bartlett-Large H	KV	0.055	0.066	0.103			
Bartlett-Med. H	N(0,1)	0.120	0.142	0.200	0.442	0.482	0.518
Bartlett-Med. H	KV	0.054	0.070	0.111			
Bartlett-Small H	N(0,1)	0.094	0.116	0.177	0.076	0.108	0.154
Bartlett-Small H	KV	0.056	0.075	0.126			
CCE-Large L	t(G-1)	0.053	0.060	0.082	0.050	0.074	0.074
CCE-Med. L	t(G-1)	0.055	0.070	0.116	0.052	0.072	0.096
CCE-Small L	t(G-1)	0.056	0.080	0.157	0.064	0.148	0.226

Note: The table reports rejection rates for 5% level tests from a Monte Carlo simulation experiment. The time series simulations are based on 30,000 simulation replications and the spatial simulations are based on 500 simulation replications. Row labels indicate which covariance matrix estimator is used. Column 2 indicates which reference distribution is used with KV corresponding to the Kiefer and Vogelsang (2005) approximation. IID and Heteroskedasticity use conventional OLS standard error and heteroskedasticity robust standard errors respectively. Rows labeled Bartlett use HAC estimators with a Bartlett kernel. Rows labeled CCE use the CCE estimator. Small, Medium, and Large denote lag truncation parameters for HAC or number of observations per group for CCE. For time series models, Small, Medium, and Large respectively denote bandwidths of 12, 20, and 38 for HAC and denote numbers of groups (G) of 12, 8, and 4 for CCE. For spatial models, Small, Medium, and Large denote bandwidths of 14, 122, and 486 for HAC and denote numbers of groups (G) of 144, 16, and 4 for CCE.

Table 2. Simulation Results from Unemployment Data.
T-test Rejection Rates for 5% Level Tests

		AR(13)		AR(1)	
	Ref. Dist.	γ = .8	γ = .4	FD	FE
IID	N(0,1)	0.476	0.430	0.142	0.683
Heteroskedasticity	N(0,1)	0.486	0.431	0.147	0.685
Cluster:					
State	t(48)	0.170	0.132	0.140	0.253
Month	t(383)	0.314	0.303	0.059	0.489
State/Month	t(48)	0.134	0.106	0.057	0.190
G4 x T3	t(11)	0.127	0.085	0.071	0.103
G4 x T6	t(23)	0.175	0.131	0.076	0.121
G4 x T32	t(127)	0.389	0.355	0.078	0.166
G2 x T3	t(5)	0.085	0.070	0.065	0.083
G2 x T6	t(11)	0.129	0.114	0.060	0.101
G2 x T32	t(63)	0.360	0.335	0.063	0.154
T3	t(2)	0.076	0.058	0.056	0.052
T6	t(5)	0.107	0.092	0.044	0.066
T32	t(31)	0.331	0.324	0.049	0.109
G4	t(3)	0.077	0.066	0.079	0.066
G2	t(1)	0.054	0.054	0.057	0.068
State x T3	t(146)	0.210	0.165	0.147	0.271
State x T6	t(493)	0.256	0.207	0.149	0.284
State x T32	t(1567)	0.484	0.435	0.146	0.344

Note: The table reports rejection rates for 5% level tests from a Monte Carlo simulation experiment with BLS unemployment data regressed on a randomly generated treatment controlling for state and month effects. All results are based on 1000 simulation replications. For the AR(1) example, we consider both fixed effects (FE) and first-differencing to remove state effects and include month dummies. For the AR(13), we use first-differencing to remove the state effects and include a full set of month dummies. The parameter γ controls the strength of cross-sectional dependence and is set to .8 in the AR(1) simulations. See text for further details about the simulation design. Rows labeled IID and Heteroskedasticity use conventional OLS and heteroskedasticity consistent standard errors respectively. The remaining rows used the CCE with different grouping schemes. "State" and "Month" use states and months as groups, respectively. "State/Month" treats observations as belonging to the same group if they belong to the same state or the same month. For the remaining groups, G2 and G4 respectively indicate partioning groups into two and four geographic regions. T3, T6, and T32 divide the time series into three 128-month periods, six 64-month periods, or 32 twelve-month periods. "G4 x T3" then indicates a group structure where observations in region one in time period one belong to the same group, observations in region two in time period one belong to the same group, etc. The sample size is N=49 and T=383.

Figure 1: Size-Adjusted Power Curve for Test Using CCE with 4 Groups and HAC with Bandwidth 38 and Kiefer-Vogelsang (2005) Reference Distribution for Time Series Simulation with $\rho = 0.8$

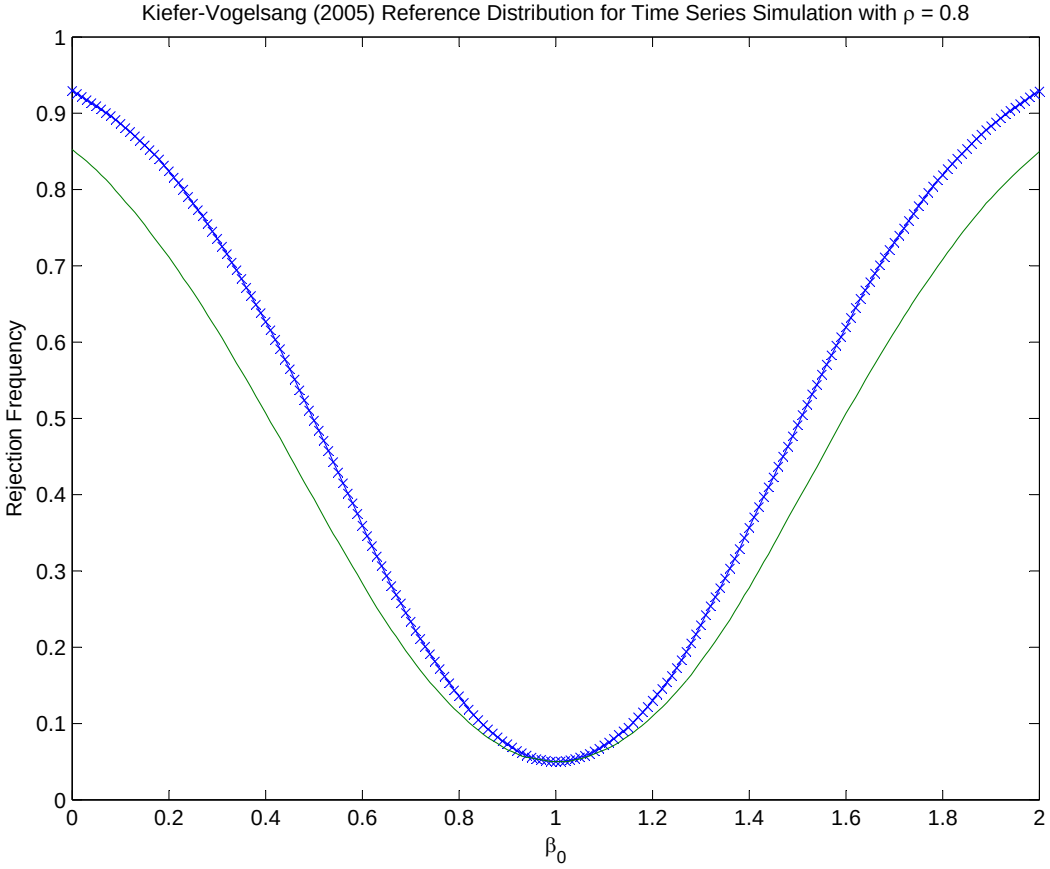

Figure 2: Power Curve for Test Using CCE with G4, T3, and G2 in Unemployment Simulation with AR(13) and $\gamma = .4$

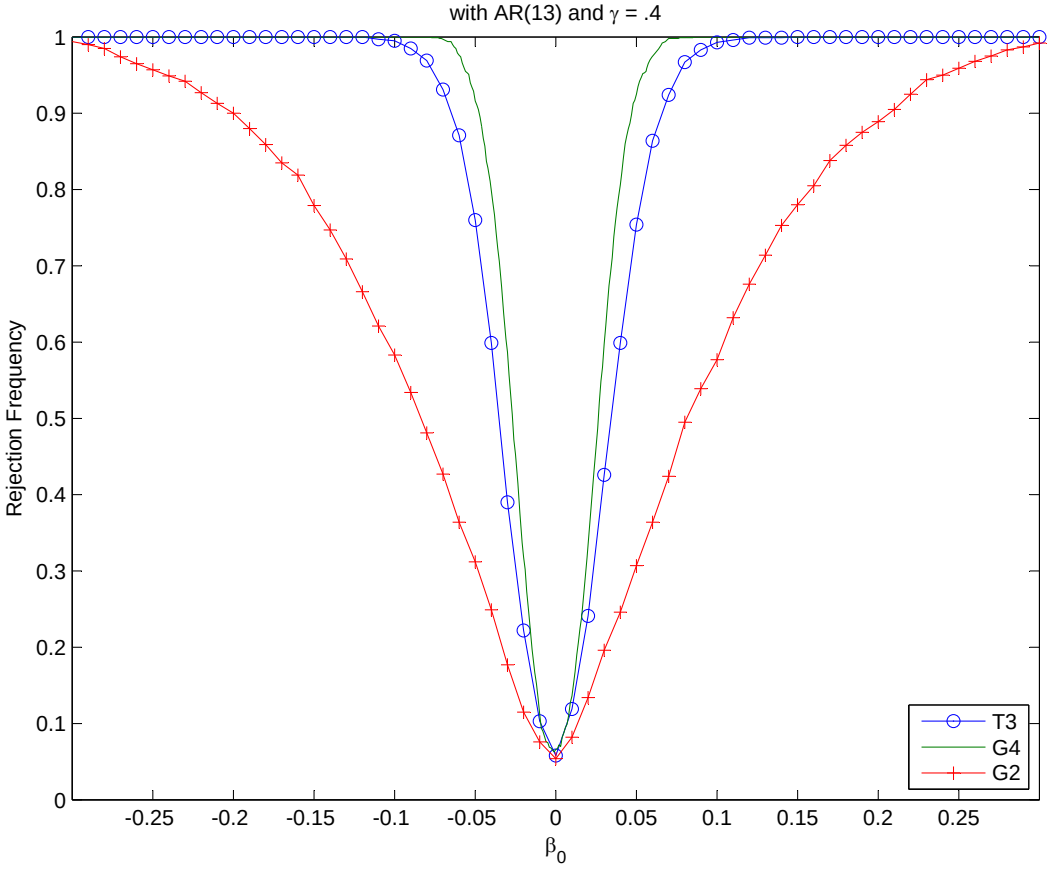

Table 3. t-test Rejection Frequencies, BCH vs IM, 5% Nominal Size
Homoskedastic Designs

		Test	Bias	RMSE	Size
OLS with Serial Correlation		BCH	-0.001	0.131	0.058
		IM	-0.001	0.134	0.052
2SLS 3 Instruments	Benchmark PI=2	BCH	-0.003	0.065	0.061
		IM	-0.013	0.068	0.062
	PI = 1	BCH	-0.017	0.130	0.068
		IM	-0.062	0.148	0.103
	PI = .5	BCH	-0.057	0.269	0.096
		IM	-0.261	0.371	0.232
2SLS 6 Instruments	PI = 2	BCH	-0.013	0.065	0.067
		IM	-0.049	0.080	0.114
	PI = 1	BCH	-0.050	0.131	0.089
		IM	-0.172	0.203	0.269
	PI = .5	BCH	-0.181	0.275	0.179
		IM	-0.444	0.469	0.585

Column labeled Size is rejection frequency across simulations. Column headings Bias and RMSE refer to the bias and root mean squared error across simulations of the numerators of the test statistics. All results based on 10,000 simulation replications with T = 100. G = 4.

Table 4. t-test Rejection Frequencies, BCH vs IM, 5% Nominal Size
"Heteroskedastic" Designs - 2SLS

			Bias	RMSE	Size
	Std Dev (Group Variances) = .064	BCH	-0.002	0.058	0.057
		IM	-0.011	0.062	0.061
PI = 2	Std Dev (Group Variances) = .22	BCH	-0.002	0.046	0.061
		IM	-0.01	0.053	0.051
	Std Dev (Group Variances) = .90	BCH	-0.001	0.03	0.071
		IM	-0.007	0.045	0.038
	Std Dev (Group Variances) = .064	BCH	-0.012	0.115	0.067
		IM	-0.05	0.133	0.095
PI = 1	Std Dev (Group Variances)= .22	BCH	-0.01	0.093	0.065
		IM	-0.041	0.113	0.077
	Std Dev (Group Variances) = .90	BCH	-0.003	0.061	0.072
		IM	-0.027	0.102	0.045
	Std Dev (Group Variances)= .064	BCH	-0.046	0.242	0.086
		IM	-0.224	0.328	0.205
PI = .5	Std Dev (Group Variances)= .22	BCH	-0.035	0.191	0.077
		IM	-0.168	0.265	0.155
	Std Dev (Group Variances)= .90	BCH	-0.019	0.122	0.075
		IM	-0.124	0.224	0.086

There are three instruments in all cases. Column labeled Size is rejection frequency across simulations. Column headings Bias and RMSE refer to the bias and root mean squared error across simulations of the numerators of the test statistics. All results based on 10,000 simulation replications with T = 100. G = 4. Std Dev(Group Variances) refers to standard deviations across groups of within-group sample variances.