

Inference with Dependent Data Using Cluster Covariance Estimators

C. Alan Bester, Timothy G. Conley, and Christian B. Hansen*

First Draft: February 2008. This Draft: January 2009.

Abstract. This paper presents a novel way to conduct inference using dependent data in time series, spatial, and panel data applications. Our method involves constructing t and Wald statistics utilizing a cluster covariance matrix estimator (CCE). We then use an approximation that takes the number of clusters/groups as fixed and the number of observations per group to be large and calculate limiting distributions of the t and Wald statistics. This approximation is analogous to ‘fixed-b’ asymptotics of Kiefer and Vogelsang (2002, 2005) (KV) for heteroskedasticity and autocorrelation consistent inference, but in our case yields standard t and F distributions where the number of groups essentially plays the role of sample size. We provide simulation evidence that demonstrates our procedure outperforms conventional inference procedures and performs well comparably to KV.

Keywords: HAC, panel, robust, spatial

JEL Codes: C12, C21, C22, C23

1. Introduction

Many economic applications involve data that cannot be modeled as independent. 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 it is true that the dependence structure in a given data set is not the object of chief interest in many economic applications, it is well understood that inference about parameters of interest, such as

* The University of Chicago, Graduate School of Business, 5807 South Woodlawn Avenue, Chicago, IL 60637, USA.

regression coefficients, may be severely distorted when one does not account for dependence in the data. 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 inference with dependent data. By far the most common is to utilize a limiting normal approximation that depends on an unknown variance-covariance matrix. This approximation is then used by simply ‘plugging-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.² 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 approximation to that employed in 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 calculate an approximate distribution for commonly-used test statistics accounting for this random limit of 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 in the limit and thus uses a ratio of limiting random variables as a reference distribution, where the numerator has the usual asymptotic normal limit for a regression parameter estimator. The resulting limit distribution for the t-statistic is pivotal but nonstandard, so critical values are obtained by

²This HAC estimator is most often a smoothed periodogram estimator but could of course be a series estimator, e.g. a flexible vector autoregression.

simulation. For many time series applications, KV provide convincing 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.³ 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). The main contribution of this paper is to provide conditions on group structure and dependence across observations such that we may derive the behavior of test statistics formed using the CCE. Our most interesting results are obtained under asymptotics that treat the number of groups as fixed and the number of observations within a group as large. We also present consistency results for the CCE when both the number of groups and their size are allowed to grow at certain rates.

Cluster covariance estimators are routinely used with data that has a group structure with independence across groups.⁴ 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. Though this is sometimes assumed for convenience, observations in different groups are generally *not* independent; for example, consider firms in the same industry in subsequent

³Bester, Conley, Hansen, and Vogelsang (2008) considers a different extension of KV to spatially dependent data. They consider spatial HAC estimators as in Conley (1996, 1999), which are a direct analog of time series HAC estimators, rather than the cluster covariance estimator considered in this paper.

⁴See 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).

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. contiguous points on the line), the overwhelming majority of points in a group will be far from other groups, and hence approximately independent of other groups provided the data are weakly dependent. The ability of the researcher to construct groups whose averages are approximately independent is the key prerequisite for our methods. 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. The idea dates to at least Bartlett (1950). What has become known as the Bartlett spectral density estimator is motivated by analogy to an average of periodogram estimates across a partition of the dataset, a procedure analogous to clustering. 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 and important 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, and they provide primitive conditions for this to be satisfied in a time series. 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.⁵

⁵We 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 Appendix B.

Our main 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. Thus our assumptions are no stronger than those routinely made with the plug-in HAC approach. Using arguments in Ibragimov and Müller (2006), we also show that standard t-statistics are conservative for hypotheses tests with size less than about 8% when the homogeneity restrictions involving unobservables are removed.

We also demonstrate that the CCE is a valid HAC estimator; that is, it is consistent under asymptotics in which the number of groups and number of observations per group go to infinity at appropriate rates. The conditions we use are analogous to the conditions used for establishing consistency of conventional time series or spatial HAC estimators with minor additional restrictions relating to group structure. Under this sequence the usual limiting normal and χ^2 approximations provide valid inference. We note that using distributional approximations based on a fixed number of groups will remain valid under this sequence since they approach the same normal and χ^2 limits with a large number of groups. Also, as illustrated in our simulation results, one will often wish to use relatively few groups to produce tests with approximately correct size. These arguments strongly suggest using the fixed-number-of-groups approximations in all cases.

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. Excellent examples of

papers in this literature are Conley (1996, 1999), Kelejian and Prucha (1999, 2001), Lee (2004, 2007a, 2007b), and Jenish and Prucha (2007). We note that analysis of spatially dependent data is not a trivial extension of results for scalar-indexed (time series). Complications arise due to such concerns as set boundaries being of large order of magnitude relative to set sizes and the number of potential neighbors of any particular point increasing rapidly with the dimension in which dependence increases. We provide formal conditions under which inference based on the CCE remains valid in very general settings.

Finally, we present simulation evidence on the performance of our estimator in time series, spatial, and panel data contexts. We provide results in a time series setting and in a cross sectional setting with spatial dependence using 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: Even modest serial or spatial dependence needs to be accounted for in order to produce reliable inference. 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 good power properties.

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, large-sample properties of the CCE itself, and the extension of our method to nonlinear models. 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) estimation of the linear model.

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 on an m -dimensional integer lattice, \mathbb{Z}^m . 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, \dots, s_N . This is routine for time series data where these indices reflect the timing of the observations. In addition, 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.⁶ 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

⁶The economics of the application often provides considerable guidance regarding the index space/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.

related/correlated and correlation patterns within sets of observations can be quite complicated with multidimensional indices.

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 $g = 1, \dots, G_N$. 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 tests of hypotheses regarding β and/or construct confidence intervals for β . We will examine the OLS estimator of β using the whole sample, which of course can be written as

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

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 \tilde{V}_N , of the variance matrix of partial sums of $x_{s_i} \varepsilon_{s_i}$, along with the usual large-sample approximation for the distribution of $\hat{\beta}_N$. Specifically, the large-sample distribution of $\hat{\beta}_N$ is

$$\sqrt{N} (\hat{\beta}_N - \beta) \xrightarrow{d} N(0, Q^{-1} V Q^{-1})$$

$$V = \lim_{N \rightarrow \infty} \text{Var} \left(\frac{1}{\sqrt{N}} \sum_{i=1}^N x_{s_i} \varepsilon_{s_i} \right)$$

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, \tilde{V}_N , of V to arrive at the approximation:

$$(2.1) \quad \hat{\beta}_N \stackrel{Approx}{\sim} N \left(\beta, \frac{1}{N} \left[\frac{1}{N} \sum_{i=1}^N x_{s_i}x'_{s_i} \right]^{-1} \tilde{V}_N \left[\frac{1}{N} \sum_{i=1}^N x_{s_i}x'_{s_i} \right]^{-1} \right)$$

Conventionally, one would use an estimator for \tilde{V}_N that is consistent for V under general forms of conditional heteroskedasticity and autocorrelation. 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 \hat{V}_{HAC} .

When the data is located at integers on the line, say $s_1 = 1, \dots, s_N = N$, spatial and discrete time series estimators for V coincide and typically are written as a weighted sum of sample autocovariances with weights 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 $W_N(h) \rightarrow 1$ for all h slowly enough for the variance of \hat{V}_{HAC} to vanish as $N \rightarrow \infty$; see, e.g., Andrews (1991). Perhaps the most popular choice for weight function $W_N(h)$ is the Bartlett kernel: an isocoles triangle that is one at $h = 0$ with a base of width $2H_N$: $W_N(h) = (1 - \frac{|h|}{H_N})^+$. The smoothing parameter H_N is assumed to grow slowly enough with the sample size for the variance of \hat{V}_{HAC} to vanish.

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:

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

where $\text{dist}(s_i, s_j)$ 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 $W_N(d) \rightarrow 1$ for all d slowly enough for the variance of \hat{V}_{HAC} to vanish as $N \rightarrow \infty$; see Conley (1999). The typical empirical approach is to choose a weight function $W_N(\cdot)$ and compute \hat{V}_{HAC} to plug into expression (2.1). The resulting distributional approximation is then used to conduct hypothesis testing and construct confidence regions.

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 \hat{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.

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).⁷ The CCE may be defined as follows:

$$\hat{V}_N \equiv \frac{1}{G_N} \sum_{g=1}^{G_N} \frac{1}{L_N} x'_g e_g e'_g x_g$$

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

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

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: It is a (spatial) HAC estimator with a discrete group-membership metric.

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 Mulainathan (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 in \mathbb{Z}^m . In Section 3.2 we demonstrate that \hat{V}_N can still be a consistent estimator of V if G_N and L_N grow at the proper rates. However, the main results in our paper focus on cases where $G_N = G$ is taken as fixed, so that \hat{V}_N converges to a random variable (i.e., is not a consistent estimator of V).

⁷A direct extension of KV to the spatial case, using spatial HAC estimators as proposed by Conley (1996, 1999), is considered in Bester, Conley, Hansen, and Vogelsang (2008).

Our method uses \hat{V}_N to form an estimator of the asymptotic variance of $\hat{\beta}$,

$$(2.2) \quad \frac{1}{N} \left[\frac{1}{N} \sum_{i=1}^N x_{s_i} x'_{s_i} \right]^{-1} \hat{V}_N \left[\frac{1}{N} \sum_{i=1}^N x_{s_i} x'_{s_i} \right]^{-1},$$

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 $L_N \rightarrow \infty$. 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 obviously be obtained in the usual fashion given these limiting results. 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.⁸⁹

Throughout the paper we will refer to a partitioning of the data into groups of contiguous 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. In m-dimensions, we impose the additional

⁸The exact scaling in Stata 10 is slightly different than ours due to the presence of a small-sample degrees of freedom correction. Specifically, $\hat{V}_{STATA} = \frac{N-1}{N-k} \frac{G}{G-1} \hat{V}_N$; see *Stata User's Guide Release 10* p. 275. Thus, scaling the STATA t-statistic by multiplying it by $\sqrt{\frac{N-1}{N-k}}$ would be equivalent to our recommended procedure. Obviously, there 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.

⁹We note that the confidence intervals reported by Stata after the use of the cluster command do use critical values from a Student-t distribution with $G - 1$ degrees of freedom.

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. For example, in the time series case with quarterly data, clustering by decade is permitted under our assumptions, but clustering by quarter (i.e., all of the first quarter observations are one group) is not. 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 aggregates 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. This result corresponds to asymptotic results for HAC estimators with smoothing parameter proportional to sample size, or “fixed-b” asymptotics, found in recent work by Kiefer and Vogelsang (2002, 2005), Phillips, Sun, and Jin (2003), and Vogelsang (2003). In our result, the number of observations per group roughly plays the role of the HAC smoothing parameter. We show that under sensible sampling conditions, inference using the CCE may still be conducted using standard t and Wald statistics even though the CCE is not consistent since these statistics follow limiting t and F distributions. For completeness, we also show that the CCE is consistent under asymptotics where the number of groups and the number of observations per group both go to infinity. This result parallels results for time series HAC as in, for example, Andrews (1991) and for spatial HAC as in, for example, Conley (1999). For reasons outlined below, we advocate the use of the fixed-G results in practice.

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.

For simplicity we will index observations on an m -dimensional integer lattice, \mathbb{Z}^m , and use the maximum coordinatewise metric $dist(s_i, s_j)$.¹⁰ Throughout, let $\mathcal{G}_{g_1}, \mathcal{G}_{g_2}$ be two disjoint sampling regions (index sets) corresponding to groups $\{g_1, g_2\} \subseteq \{1, \dots, G\}$ 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 $N \rightarrow \infty$.
- (ii) As $N \rightarrow \infty$, $L_N \rightarrow \infty$ and G is fixed.
- (iii) $\{x_s, \varepsilon_s\}$ is α -mixing¹¹ 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 $\eta > 0$.

¹⁰The maximum coordinatewise distance metric is defined as $dist(s_i, s_j) = \max_{l \in \{1, \dots, m\}} |s_i(l) - s_j(l)|$ 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.

¹¹We 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}_Λ be the σ -algebra generated by a given random field $\psi_{s_m}, s_m \in \Lambda$ 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 Λ_2 . For our results, we use the maximum coordinate-wise distance metric. The mixing coefficient is then

- $\sup_s \mathbb{E}|\varepsilon_s|^{2r} < \infty$ and $\sup_s \mathbb{E}|x_{sh}|^{2r} < \infty$ for $r > 2 + \delta$ where x_{sh} is the h^{th} element of vector x_s . $\mathbb{E}[\frac{1}{L_N}x'_g x_g]$ is uniformly positive definite with constant limit Q_g for all $g = 1, \dots, G$.
- (iv) $\mathbb{E}[x_s \varepsilon_s] = 0$. $V_{Ng} = \text{var}[\frac{1}{\sqrt{L_N}}x'_g \varepsilon_g]$ is uniformly positive definite with constant limit Ω_g for all $g = 1, \dots, G$.

Part (i) simply ensures that indexing in m -dimensions is required, i.e. that indexing in a lower dimension space is not adequate. Part (ii) restates our asymptotic sequence definition that there are a fixed number of groups whose size is increasing. The key part of (iii) is the mixing and moment conditions. The mixing conditions could of course be replaced with an analogous set of conditions using another definition of mixing or another notion of weak dependence to limit the dependence in the data to a level that permits a central limit theorem.

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 $\text{dist}(\cdot)$.*
- (iv) *For all g , $|\partial\mathcal{G}| < CL_N^{\frac{m-1}{m}}$.*

Part (i) of Assumption 2 could be relaxed to allow an asymptotically negligible amount of overlap across groups. Part (ii) of Assumption 2 assumes a common group size.¹² Part (iii) of Assumption 2 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 that resemble $\overline{\alpha_{k,l}(j)} \equiv \sup\{|P(A \cap B) - P(A)P(B)|\}$, $A \in \mathcal{F}_{\Lambda_1}$, $B \in \mathcal{F}_{\Lambda_2}$, and $|\Lambda_1| \leq k$, $|\Lambda_2| \leq l$, $\Upsilon(\Lambda_1, \Lambda_2) \geq j$. Mixing requires that $\alpha_{k,l}(j)$ converges to zero as $j \rightarrow \infty$.

¹²We 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 $L_{g_1}/L_{g_2} \rightarrow 1$ for all g_1 and g_2 . See also the discussion about weighting following the statement of Proposition 1.

a collection of regular polyhedra growing to cover the space. For example, in the special case of a time series ($m = 1$), (iii)-(iv) requires essentially that groups are ‘blocks’ on the line and rules out groups consisting of every k th observation. The boundary condition in (iv) is the key element used for our results.

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

$$(i) \quad \frac{1}{L_N} \begin{pmatrix} x'_1 x_1 \\ \vdots \\ x'_G x_G \end{pmatrix} \xrightarrow{p} \begin{pmatrix} Q_1 \\ \vdots \\ Q_G \end{pmatrix} \text{ and}$$

$$(ii) \quad \frac{1}{\sqrt{L_N}} \begin{pmatrix} x'_1 \varepsilon_1 \\ \vdots \\ x'_G \varepsilon_G \end{pmatrix} \xrightarrow{d} N \left(0, \begin{pmatrix} \Omega_1 & & 0 \\ & \ddots & \\ 0 & & \Omega_G \end{pmatrix} \right)$$

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

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 essentially the same as needed 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 $L_N \rightarrow \infty$, 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 matrices of cross-products of x converge to the same limit within each group and that the asymptotic variance of the score within each group is constant. These conditions are implied by covariance stationarity of the individual observations but may also be satisfied even if covariance stationarity is violated.

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

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

or, when $q = 1$,

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

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

Proposition 1. Suppose $\{\mathcal{G}_g\}$ is defined such that $L_N \rightarrow \infty$ and G is fixed as $N \rightarrow \infty$ 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 ,

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

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

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

$$P(|t_{G-1}^*| > c_{G-1}(\alpha)) \leq \alpha$$

for $c_{G-1}(\alpha)$ the usual critical value for an α -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. In the time series case, the reference distributions under homogeneity are simpler to work with than the KV reference distributions which also rely on homogeneity. Moreover, our results apply immediately to general spatial contexts, including panel data and cross-sectional dependence.

Proposition 1, part (i) imposes essentially no homogeneity and implicitly allows for group sizes that are not asymptotically equivalent. Without further restrictions, we see that the CCE converges to a limiting random variable and that usual test statistics formed using this covariance matrix estimator take the form of ratios of random variables. The limiting distributions of the test statistics

are neither standard nor pivotal though in principle 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 $G_N \rightarrow \infty$, 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 $G_N \rightarrow \infty$.¹³ 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} .

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

¹³With 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 $L_N \rightarrow \infty$ and then $G_N \rightarrow \infty$. 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 following section under asymptotics where $\{L_N, G_N\} \rightarrow \infty$ jointly.

converge to scaled t- and F-distributions.¹⁴ 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.

It is worth noting that the assumption of common group sizes is purely for simplicity in Part (i) of Proposition 1, as we have placed no structure on Q_g or Ω_g across groups. This is not the case for Parts (ii) and (iii) of Proposition 1, in particular because Assumption 3 is probably not reasonable for asymptotically heterogeneous group sizes. It could be shown that a version of Part (ii) of Proposition 1 holds for weighted least squares with heterogeneous group sizes if Assumption 2.(ii) and Assumption 3 are replaced with the following:

Assumption 5. For all g , $|\mathcal{G}_g| = L_{g,N}$, and $L_{g,N}/\bar{L}_N \rightarrow \rho_g$ where $\bar{L}_N = \frac{1}{G} \sum_g L_{g,N}$.

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

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 $\sqrt{\bar{L}_N/L_{g,N}}$. Defining \hat{t}_w as above using the WLS estimates in place of $\hat{\beta}_N$, \hat{Q} , and \hat{V}_N , Proposition 1 Part (ii) obtains immediately for \hat{t}_w under the Assumptions 1, 5, and 6.

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

¹⁴ We thank Jim Stock and Mark Watson for pointing out the F-statistic result. See also Stock and Watson (2008).

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.

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 average of the group-level estimates in the numerator and the standard deviation of the estimates across groups in the denominator.¹⁵ Under high level assumptions, 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 use the same intermediate result: 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). Of course, Lemma 1 is crucial for our results as well and a main contribution of our paper is the provision of primitive conditions (in the form of Assumptions 1 and 2) that are sufficient for the conclusion of Lemma 1 to hold in sampling environments where dependence is indexed in m-dimensions.

Since we share the intermediate result of Lemma 1's conclusion with IM, our primitive conditions in Assumptions 1 and 2 imply that the IM results are also valid with empirically relevant forms of m-dimensionally indexed dependence. This considerably increases the set of applications where the IM approach is formally justified (in terms of primitive conditions). This is particularly

¹⁵Our 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.

useful in applications with pronounced heterogeneity in regressor variances across groups where our homogeneity restrictions would not apply but the IM approach remains valid.¹⁶

Our approach and that of IM are best thought of as complements as there are clearly scenarios where one would be preferred over the other. The most obvious distinction is that our methods extend immediately to joint hypothesis tests while IM methods apply only to hypotheses about scalar parameters. 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 some simulation results in Appendix B 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 should therefore both prove useful to applied researchers.

3.2. $G \rightarrow \infty$ Asymptotics

We believe that the results from the previous section suggest that appropriately normalized t- and F-statistics should always be used for conducting inference when using the CCE in practice. In this section, we provide a standard consistency result for the CCE estimator under weak dependence

¹⁶As we mention above, our approach can be generalized to allow for some types of heterogeneity by adopting a weighting scheme across groups.

when $\{G_N, L_N\} \rightarrow \infty$ jointly. As in the fixed- G case, L_N plays the role of the usual smoothing parameter or lag truncation parameter for a HAC estimator, and we obtain consistency under conditions on the rate of growth of L_N that correspond to the usual rate conditions on the smoothing parameter for a conventional HAC estimator. This analysis serves to unify the CCE and HAC estimators and technically complete the analysis. We present the results for time series and spatial processes separately as the rates obtained under the two cases are substantially different.

In stating the result, we will make use of the following assumptions. We use Assumption 7 to establish the consistency of the CCE in the time series context or when space is indexed on a line, and we use Assumption 8 to establish consistency under general spatial dependence.

Assumption 7 (Time Series/Locations on Line).

- (i) As $N \rightarrow \infty$, $L_N \rightarrow \infty$ such that $\frac{L_N}{N} \rightarrow 0$.
- (ii) $\{x_s, \varepsilon_s\}$ is an α -mixing sequence that satisfies $\sum_{j=1}^{\infty} j^2 \alpha(j)^{\delta/(4+\delta)} < \infty$ for some $\delta > 0$, and $\sup_s E|\varepsilon_s|^{8+2\delta} < \infty$ and $\sup_s E|x_{sh}|^{8+2\delta} < \infty$ where x_{sh} is the h^{th} element of vector x_s . $E[\frac{1}{N} \sum_s x_s x_s']$ is uniformly positive definite with constant limit \mathbf{Q} .
- (iii) $E[x_s \varepsilon_s] = 0$. $V_N = \text{var}[\frac{1}{\sqrt{N}} \sum_s x_s \varepsilon_s]$ is uniformly positive definite with constant limit V .

These conditions are quite standard in the HAC literature; see, for example, Andrews (1991). Under these conditions, consistency and asymptotic normality of the least squares estimator are easily established, as is consistency of the infeasible HAC estimator which uses smoothing parameter L_N and the actual ε_s 's. Condition (i) imposes a condition on the rate of growth of the number of observations per group and thus implicitly defines a rate of growth on the number of groups, G_N . We note that this condition is quite mild and will be satisfied as long as G_N grows at any rate since all that is required is that $\frac{L_N}{N} = \frac{1}{G_N} \rightarrow 0$.

Assumption 8 (Locations in m Dimensions).

- (i) The sample region grows uniformly in m non-opposing directions as $N \rightarrow \infty$.
- (ii) $\text{Diam}(\mathcal{G}_g) \leq CL^{1/m}$ for all g where $\text{Diam}(\mathcal{G}) = \max\{\text{dist}(s_i, s_j) | i, j \in \mathcal{G}\}$, and the maximum number of groups within distance $CL^{1/m}$ of any particular group g is bounded for all g .
- (iii) As $N \rightarrow \infty$, $L_N \rightarrow \infty$ such that $\frac{L_N^3}{N} \rightarrow 0$.
- (iv) $\{x_s, \varepsilon_s\}$ is a α -mixing and satisfies $\alpha_{\infty, \infty}(j)^{\delta/(2+\delta)} = O(j^{-2m-\eta})$ for some $\delta > 0$ and some $\eta > 0$, and $\sup_s E|\varepsilon_s|^{8+2\delta} < \infty$ and $\sup_s E|x_{sh}|^{8+2\delta} < \infty$ where x_{sh} is the h^{th} element of vector x_s . $E[\frac{1}{N} \sum_s x_s x_s']$ is uniformly positive definite with constant limit \mathbf{Q} .
- (v) $E[x_s \varepsilon_s] = 0$. $V_N = \text{var}[\frac{1}{\sqrt{N}} \sum_s x_s \varepsilon_s]$ is uniformly positive definite with constant limit V .

Conditions (i) and (iii)-(v) of Assumption 8 are the same as those used in Conley (1999) and are sufficient to guarantee the consistency and asymptotic normality of the OLS estimator and to guarantee the consistency of HAC estimators including the infeasible spatial HAC estimator that uses the true values of the ε_s 's. Condition (i) gives content to the notion of sampling in m dimensional space, since the problem would effectively be of lower than m dimensions if the sample region increased in less than m non-opposing dimensions. One might wish to relax the condition that the rate of increase is uniform, but we leave this to future work. Condition (ii) imposes a restriction on the choice of groups in constructing the CCE. This condition will be satisfied, for example, when the groups are constructed as regular polyhedra. It rules out cases where one constructs groups with very long, skinny parts where there can be arbitrarily many groups within a particular smoothing parameter. Condition (iii) then gives the rate of growth on the smoothing parameter relative to the sample size. Again this condition imposes a rate condition on the number of groups, implying that $\frac{L_N^2}{G_N} \rightarrow 0$. That is, we see that consistency will obtain only when the number of groups is growing quite quickly relative to the number of observations per group. Since reasonable levels of dependence imply groups need to be large, this condition on the number of groups indicates this approximation will only be useful when N is extremely large. We also note that these rate conditions are substantially different than those obtained in the time series case. It

may be interesting to see if the rate conditions could be relaxed, but doing so is beyond the scope of the present paper.

Under Assumption 7 or 8, we can state the following result.

Proposition 2. *If Assumption 7 is satisfied and data are indexed on a line (e.g. a time series) or Assumption 8 is satisfied, $\widehat{V}_N \xrightarrow{p} V$ where $\sqrt{N}(\widehat{\beta} - \beta) \xrightarrow{d} \mathbf{Q}^{-1}N(0, V)$.*

Proposition 2 verifies that the CCE, \widehat{V}_N , is consistent under an asymptotic sequence where the number of groups G_N goes to infinity with the sample size. This result, along with asymptotic normality of the OLS estimator, provides an obvious approach to inference using the CCE when data are weakly dependent using the usual asymptotic normal and χ^2 approximations. As discussed in the previous section, this seems unlikely to perform as well as the fixed-G results. It is interesting that the result is obtained under the same conditions as would be used to show the consistency of conventional HAC or spatial HAC estimators, again illustrating that the CCE may be thought of as a HAC estimator with a particular kernel.

3.3. Nonlinear Models

The results from the previous sections will hold for nonlinear models under appropriate modification of regularity conditions. In this section, we provide a sketch of the requisite modifications for m-estimators for our fixed- G result. Similar modifications could be made for the case where $G_N \rightarrow \infty$.

Suppose that

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

where $\lim_{N \rightarrow \infty} \frac{1}{N} \sum_i E[f(z; \theta)]$ 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 $V = \lim_{N \rightarrow \infty} \text{Var} \left[\frac{1}{\sqrt{N}} \sum_i \frac{\partial}{\partial \theta} f(z_{s_i}; \theta_0) \right]$ and $\Gamma = \lim_{N \rightarrow \infty} E \left[\frac{1}{N} \sum_i \frac{\partial^2}{\partial \theta \partial \theta'} f(z_{s_i}; \theta_0) \right]$

under standard regularity conditions; see, for example, Wooldridge (1994) in the time series case and Jenish and Prucha (2007) in the spatial case.¹⁷

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 $g = 1, \dots, G$. 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

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

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

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

¹⁷Jenish 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).

$$+ \Gamma_g(\bar{\theta}) \left(\left[\sum_{r=1}^G \Gamma_r(\bar{\theta}) \right]^{-1} \sum_{r=1}^G D_r(\theta_0) \right) \left(\left[\sum_{r=1}^G \Gamma_r(\bar{\theta}) \right]^{-1} \sum_{r=1}^G D_r(\theta_0) \right)' \Gamma_g(\bar{\theta}) \Big].$$

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

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

where W is block diagonal with off-diagonal blocks equal to matrices of zeros and diagonal blocks equal to Ω_g where $\Omega_g = \lim_{L_N \rightarrow \infty} \text{Var} \left[\frac{1}{\sqrt{L_N}} D_g(\theta_0) \right]$ and that $\sup_{\theta \in \Theta} \left\| \frac{1}{L_N} \Gamma_g(\theta) - \Gamma_g^*(\theta) \right\| \xrightarrow{p} 0$ where $\Gamma_g^*(\theta_0)$ is nonsingular for all $g = 1, \dots, G_N$. 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.¹⁸ The second condition is a uniform convergence condition for the Hessian matrix for which a variety of primitive conditions can be found, e.g. Jenish and Prucha (2007) or Wooldridge (1994).

4. Simulation Examples

The main contribution of the previous section is in providing a limiting result under an asymptotic sequence when the number of groups remains fixed which corresponds to “fixed-b” asymptotic approximations for conventional HAC estimators. Under this asymptotics, we show that standard test-statistics follow asymptotic t - or F - 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

¹⁸Additional 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 6.3.

and state-level fixed effects are included. In all of our simulations, we consider inference about a slope coefficient from a linear regression model with point estimates obtained using OLS.¹⁹

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 $N = 1296$.²⁰

¹⁹We 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. To conserve space in the main text, these simulations are collected in Appendix B.

²⁰We draw u_s and v_s on a 40×40 lattice to generate the 36×36 lattice of x_s and ε_s .

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 4, 8, and 12 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 4, 8, and 16 for HAC and denote numbers of groups of 144, 16, and 4 for CCE.²¹

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

²¹We chose these truncation parameters for the Bartlett kernels by taking the sample size and dividing by two times the number of groups used in constructing the CCE for the time series simulation and by taking approximately the square root of the total sample size divided by the number of groups used in constructing the CCE in the spatial simulation.

of 5%, we see that the HAC tests reject 10.5% of the time with $\rho = .5$ and 18.4% of the time with $\rho = .8$ compared to 5.9% 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.

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 1.²² 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 16 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 and “comparable” smoothing parameter in this figure. We note that the power loss is much smaller across the remaining designs. The gain is the ease in computing the CCE as well as in obtaining a simple reference distribution.

²²We 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.²³ 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 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. Thus, to estimate β , we regress $\log(y_{st})$ on x_{st} and a full set of state and time dummies. We note that this is a simple but fairly standard specification in applied research and that, with unemployment rates on the left hand side, it is similar to models considered in Shimer (2001) and Foote (2007).

We use two different processes to generate x_{st} , one which generates a continuous treatment and one which generates a binary treatment. The continuous treatment is meant to represent a treatment such as the log of the youth employment share as considered in Shimer (2001) and Foote (2007) while the binary treatment is meant to represent policy, such as the presence or absence of a law. In both cases, we generate the treatment to be both spatially and intertemporally correlated.²⁴

²³We use seasonally unadjusted monthly state-level unemployment rates from the BLS available at <ftp://ftp.bls.gov/pub/time.series/la/>.

²⁴If the treatment were spatially (intertemporally) uncorrelated, $x_{st}\varepsilon_{st}$ would also be spatially (intertemporally) uncorrelated regardless of the correlation structure of ε_{st} . We note that there is a high degree of

For the continuous treatment, we generate x_{st} as

$$x_{st} = \sigma_1 \left(u_{st} + \gamma \sum_{d(s,r)=1} u_{rt} \right) \text{ where}$$

$$u_{st} = \mu + \rho u_{s(t-1)} + v_{st},$$

$$v_{st} \sim N(0, \sigma_2^2),$$

$d(s, r)$ is one for adjacent states s and r and zero otherwise, and ρ and γ respectively control the strength of the intertemporal and spatial correlation and are varied in the simulations. We choose μ , σ_1 , and σ_2 to make the mean and variance of x_{st} similar to the mean and variance of $\log(y_{st})$,²⁵

When x_{st} is binary, we generate a random time, τ_s , for each state to adopt the treatment:

$$\tau_s = \left(u_s + \gamma \sum_{d(s,r)=1} u_r \right) / \left(1 + \sum_{d(s,r)=1} \gamma \right)$$

where u_s is a discrete uniform distribution with support $\{1, \dots, 384\}$ and $d(s, r)$ is defined above. We then define $x_{st} = \mathbf{1}(t > \tau_s)$ where $\mathbf{1}(\cdot)$ is equal to one if the event in the parentheses is true and zero otherwise. In this case, the treatment is highly serially dependent and γ again controls the degree of spatial correlation.

We report simulation results for the continuous treatment design in Table 2 and the binary treatment design in Table 3. In all cases, we report rejection frequencies for 5% level tests of the hypothesis that $\beta = 0$. For the continuous treatment, we report results with γ of 0.2 or 0.8 and ρ of 0.4 or 0.8. For the binary treatment, we report results for $\gamma \in \{0, 0.2, 0.4, 0.8\}$. 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

both spatial and intertemporal correlation in the log of the unemployment rate even after controlling for state and time effects.

²⁵We used $\sigma_1 = \sigma_2 = .2$ and $\mu = .4$.

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.²⁶ 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. Based on the simulation results from the previous section, we did not compute HAC standard errors as they are computationally more burdensome than the CCE and had substantial size distortions when the KV reference distribution was not used. For all simulations, we use the full sample with 49 states and 384 time periods, and all results are based on 1000 simulation replications.²⁷

Looking first at Table 2 which has results for the continuous treatment, we see 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) perform uniformly poorly across the designs considered. With a moderate value of either γ or ρ , we see that the grouping strategies that use a large number of groups fare quite poorly. On the other hand, the conservative grouping

²⁶For 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.

²⁷We have run simulations using various subsamples which produce qualitatively similar results to those reported here with the caveat that as the time series dimension decreases it becomes increasingly obvious, as intuition would suggest, that there is no sense in entertaining groups splitting on the time dimension. We also considered values of γ equal to .4 and 0 in the continuous treatment case. The results for $\gamma = .4$ are intermediate to those reported, and unsurprising, simply clustering by state works very well when $\gamma = 0$.

strategies; T3, T6, G2, and G4; appear to perform well across all the values of γ and ρ . We also see that when γ and ρ are moderate, a variety of grouping strategies produce tests with size reasonably close to the nominal level.

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 convex in the degrees of freedom, there will be rapidly diminishing returns to increasing the degrees of freedom. Figure 2 plots power curves for T3, G4, and $G4 \times T3$. In the figure, the solid curve plots power for $G4 \times T3$, the crossed line plots power for G4, 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 T3 to G4, sizes are similar, but the power from G4 is substantially higher than that of T3. We also see that $G4 \times T3$ is substantially more powerful than either T3 or G4, rejecting a hypothesis that the coefficient is -0.2 over 80% of the time compared to approximately 40% of the time for T3, at the cost of a slight size distortion.

For the binary treatment case, the only strategies which reliably produce reasonably accurate testing results are G2 and G4. This result is unsurprising since both log state-level unemployment rates and the randomly generated binary treatment are strongly persistent even after taking out an arbitrary national time trend. Because of this, one is better off, in terms of size of hypothesis tests, by treating the problem as a large vector cross-sectional problem and allowing essentially arbitrary inter-temporal correlation. This example strongly illustrates that one needs to carefully consider both cross-sectional and inter-temporal correlation when conducting inference. Despite the 18,816 state-month observations, our results suggest one is faced with drawing inference from about four “independent observations” to get test size and confidence interval coverage close to the

nominal level due to the strong persistence in the time series and the moderate dependence in the cross-section.

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 approach which many people advocate 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 that is not highly misleading 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.

Overall, 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.

5. Conclusion

In this paper, we have considered the use of the clustered covariance matrix estimator (CCE) for performing inference about regression parameters when data are weakly dependent. We allow for general forms of dependence and our results apply immediately to 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.

In our main results, we consider an asymptotic sequence in which the number of groups is fixed and the number of observations per group goes to infinity. Under this sequence, we show that the CCE is not consistent but converges in distribution to a nondegenerate random variable. We then consider testing hypotheses using standard t and Wald tests based on the CCE. In this case, these test statistics do not converge to the usual normal and χ^2 limits, but 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 “fixed-b” asymptotics where the HAC smoothing parameter is allowed to be proportional to the total sample size. KV obtain asymptotic reference distributions that are nonstandard ratios of random variables, so critical values must be simulated. 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. 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.

In a secondary result, we consider a sequence where the number of groups used to form the CCE and the number of observations per group both increase with the total sample size. Under this sequence, we show that the CCE is consistent for the underlying covariance matrix of an estimator. This result is analogous to usual asymptotic results for HAC estimators that rely on a smoothing parameter parameter that increases slowly with the sample size.

An important question that deserves more attention is smoothing parameter selection which corresponds to choice of groups in our context. We note that choice of groups may be easier to understand than choice of smoothing parameter via spectral representation. 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 \widehat{V} :

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

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 \widehat{V} under Assumptions 1-2, 1-3, and 1-4 respectively:

$$\begin{aligned} V_A &= \frac{1}{G} \sum_g [\Lambda_g B_g B'_g \Lambda'_g - Q_g \mathbf{Q}^{-1} \mathbf{S} B'_g \Lambda'_g - \Lambda_g B_g \mathbf{S} \mathbf{Q}^{-1} Q_g + Q_g \mathbf{Q}^{-1} \mathbf{S} \mathbf{S}' \mathbf{Q}^{-1} Q_g] \\ V_B &= \frac{1}{G} \sum_g \left[\Lambda_g B_g B'_g \Lambda'_g - \frac{1}{G} \mathbf{S} B'_g \Lambda'_g - \frac{1}{G} B_g \Lambda_g \mathbf{S} + \frac{1}{G^2} \mathbf{S} \mathbf{S}' \right] \\ V_C &= \frac{1}{G} \Lambda \left[\sum_g B_g B'_g - \frac{1}{G} \left(\sum_g B_g \right) \left(\sum_g B'_g \right) \right] \Lambda'. \end{aligned}$$

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

$$\begin{aligned} \sqrt{N} \left(R\hat{\beta} - r \right) &\xrightarrow{d} \sqrt{GR}Q^{-1} \frac{1}{G} \sum_g \Lambda_g B_g \\ R\hat{Q}^{-1}\hat{V}\hat{Q}^{-1}R' &\xrightarrow{d} RQ^{-1}V_BQ^{-1}R' \end{aligned}$$

We can write the RHS of the second line as

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

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

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

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

$$\begin{aligned} \sqrt{N} \left(R\hat{\beta} - r \right) &\xrightarrow{d} \sqrt{GR}Q^{-1} \Lambda \left(\frac{1}{G} \sum_g B_g \right) \\ R\hat{Q}^{-1}\hat{V}\hat{Q}^{-1}R' &\xrightarrow{d} RQ^{-1}V_CQ^{-1}R' \end{aligned}$$

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 $\sum_g B_{1,g}^2 - B_{1,G}^2 \sim \chi_{G-1}^2$ 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 Proposition 2

Let $\tilde{V}_{HAC,L}$ denote the infeasible HAC estimator that uses the true ε_s 's, a uniform kernel, and smoothing parameter large enough to span any of the groups. In the time series case, this corresponds to a smoothing parameter of L , and under Assumption 8.(i), it corresponds to a smoothing parameter of $CL^{1/m}$ in each spatial dimension for some constant C . Under Assumption 7, we also have $\tilde{V}_{HAC,L} \xrightarrow{p} V$ from Andrews (1991) in the time series case, and under Assumption 8, we have $\tilde{V}_{HAC,L} \xrightarrow{p} V$ from Conley (1999).²⁸ The proof then proceeds by showing that $\hat{V} - \tilde{V}_{HAC,L} = o_p(1)$ from which $\hat{V} - V \xrightarrow{p} 0$ follows immediately. We demonstrate the result for the spatial case below. The time series case is similar but simpler and so is omitted.

First note that $\hat{V} - \tilde{V}_{HAC,L}$ can be written as

$$\hat{V} - \tilde{V}_{HAC,L} = \left(\frac{1}{N} \sum_{g=1}^G x'_g \varepsilon_g \varepsilon'_g x_g - \tilde{V}_{HAC,L} \right) - R_1 - R'_1 + R_2$$

²⁸The results in Conley (1999) assume stationarity and are worked out only for $m = 2$. These results could easily be modified under our assumptions without requiring stationarity using the central limit results of Jenish and Prucha (2007) and are also straightforward to modify under our assumptions to m -dimensional indexing.

where

$$R_1 = \frac{1}{N} \sum_{g=1}^G x'_g x_g (\widehat{\beta} - \beta) \varepsilon'_g x_g \quad \text{and} \quad R_2 = \frac{1}{N} \sum_{g=1}^G x'_g x_g (\widehat{\beta} - \beta) (\widehat{\beta} - \beta)' x'_g x_g.$$

We start by considering the first term $\frac{1}{N} \sum_{g=1}^G x'_g \varepsilon_g \varepsilon'_g x_g - \widetilde{V}_{HAC,L}$. Let $\mathcal{G}(s)$ denote the index set for the group to which the observation with index s belongs. $\frac{1}{N} \sum_{g=1}^G x'_g \varepsilon_g \varepsilon'_g x_g$ can be written as

$$\frac{1}{N} \sum_{g=1}^G x'_g \varepsilon_g \varepsilon'_g x_g = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \mathbf{1}[\mathcal{G}(s_i) = \mathcal{G}(s_j)] (x_{s_i} \varepsilon_{s_i} x'_{s_j} \varepsilon_{s_j})$$

and $\widetilde{V}_{HAC,L}$ can be written as

$$\widetilde{V}_{HAC,L} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \mathbf{1}[\text{dist}(s_i, s_j) < CL^{1/m}] (x_{s_i} \varepsilon_{s_i} x'_{s_j} \varepsilon_{s_j})$$

It follows that the first term is given by

$$\frac{1}{N} \sum_{g=1}^G x'_g \varepsilon_g \varepsilon'_g x_g - \widetilde{V}_{HAC,L} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \mathbf{1}[\text{dist}(s_i, s_j) < CL^{1/m}, \mathcal{G}(s_i) \neq \mathcal{G}(s_j)] (x_{s_i} \varepsilon_{s_i} x'_{s_j} \varepsilon_{s_j})$$

We now consider a generic element of the first term,

$$(6.1) \quad B_N^* = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \mathbf{1}[\text{dist}(s_i, s_j) < CL^{1/m}, \mathcal{G}(s_i) \neq \mathcal{G}(s_j)] (x_{s_i}^* \varepsilon_{s_i} x_{s_j}^* \varepsilon_{s_j} - \mathbb{E}[x_{s_i}^* \varepsilon_{s_i} x_{s_j}^* \varepsilon_{s_j}])$$

$$(6.2) \quad + \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \mathbf{1}[\text{dist}(s_i, s_j) < CL^{1/m}, \mathcal{G}(s_i) \neq \mathcal{G}(s_j)] \mathbb{E}[x_{s_i}^* \varepsilon_{s_i} x_{s_j}^* \varepsilon_{s_j}]$$

where the stars refer to an arbitrary element of each vector. The same argument used in Appendix 6.3 below can be used to show that we can bound the maximum number of d^{th} order neighbors from any given point within any given group, say g , that lie within some other group, say h , by $C(m)L^{(m-1)/m}d$ where m is the dimension of the index set. Under the condition on the number of neighboring groups given in Assumption 8.(ii), we can use a standard mixing inequality, for example Lemma 1 of Jenish and Prucha (2007) or Bolthausen (1982), to bound the second term in B_N^* given in (6.2) using an argument similar to that in Appendix 6.3 by $\frac{C}{N} \sum_{d=1}^{\infty} dGL^{(m-1)/m} \alpha_{1,1}(d)^{\delta/(2+\delta)} =$

$O(L^{-1/m})$ under the mixing and moment conditions of Assumption 8. It follows that

$$B_N^* = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \mathbf{1}[\text{dist}(s_i, s_j) < CL^{1/m}, \mathcal{G}(s_i) \neq \mathcal{G}(s_j)] (x_{s_i}^* \varepsilon_{s_i} x_{s_j}^* \varepsilon_{s_j} - \mathbb{E}[x_{s_i}^* \varepsilon_{s_i} x_{s_j}^* \varepsilon_{s_j}]) \\ + O(L^{-1/m}).$$

Using coordinate indexing and defining the lattice as $(1, \dots, M_1) \times \dots \times (1, \dots, M_m)$, we can write the first term in B_N^* as

$$\tilde{B}_N = \frac{1}{N} \sum_{i_1}^{M_1} \dots \sum_{i_m}^{M_m} Z_{i_1, \dots, i_m}$$

where

$$Z_{i_1, \dots, i_m} = \sum_{j_1: |j_1 - i_1| < CL^{1/m}} \dots \sum_{j_m: |j_m - i_m| < CL^{1/m}} \mathbf{1}_{i_1, \dots, i_m} \mathbf{1}_{j_1, \dots, j_m} \mathbf{1}_{\mathcal{G}(i_1, \dots, i_m) \neq \mathcal{G}(j_1, \dots, j_m)} z_{i_1, \dots, i_m, j_1, \dots, j_m},$$

$\mathbf{1}_{i_1, \dots, i_m}$ is an indicator which is one if the coordinate $s_i = (i_1, \dots, i_m)$ occurs in the sample, $\mathbf{1}_{\mathcal{G}(i_1, \dots, i_m) \neq \mathcal{G}(j_1, \dots, j_m)}$ is an indicator which is one if observations with indices $s_i = (i_1, \dots, i_m)$ and $s_j = (j_1, \dots, j_m)$ do not belong to the same group, and $z_{i_1, \dots, i_m, j_1, \dots, j_m} = x_{s_i}^* \varepsilon_{s_i} x_{s_j}^* \varepsilon_{s_j} - \mathbb{E}[x_{s_i}^* \varepsilon_{s_i} x_{s_j}^* \varepsilon_{s_j}]$.

We then have

$$\mathbb{E}|\tilde{B}_N|^2 \leq \frac{1}{N^2} \sum_{i_1}^{M_1} \dots \sum_{i_m}^{M_m} \sum_{k_1}^{M_1} \dots \sum_{k_m}^{M_m} |\mathbb{E}[Z_{i_1, \dots, i_m} Z_{k_1, \dots, k_m}]| \\ (6.3) \quad = \frac{1}{N^2} \sum_{i_1}^{M_1} \dots \sum_{i_m}^{M_m} \sum_{k_1: |k_1 - i_1| \leq 2CL^{1/m}}^{M_1} \dots \sum_{k_m: |k_m - i_m| \leq 2CL^{1/m}}^{M_m} |\mathbb{E}[Z_{i_1, \dots, i_m} Z_{k_1, \dots, k_m}]|$$

$$(6.4) \quad + \frac{1}{N^2} \sum_{i_1}^{M_1} \dots \sum_{i_m}^{M_m} \sum_{|k_1 - i_1| > 2CL^{1/m} \text{ or } \dots \text{ or } |k_m - i_m| > 2CL^{1/m}} |\mathbb{E}[Z_{i_1, \dots, i_m} Z_{k_1, \dots, k_m}]|.$$

Considering (6.3) first, we have that there are at most $(4CL^{1/m} + 1)^m$ points within $2CL^{1/m}$ of any given point. Under the conditions in Assumption 8.(iv), we also have that $|\mathbb{E}[Z_{i_1, \dots, i_m} Z_{k_1, \dots, k_m}]| \leq CL^2$ for some finite positive constant C . Therefore, we have

$$\frac{1}{N^2} \sum_{i_1}^{M_1} \dots \sum_{i_m}^{M_m} \sum_{k_1: |k_1 - i_1| \leq 2CL^{1/m}}^{M_1} \dots \sum_{k_m: |k_m - i_m| \leq 2CL^{1/m}}^{M_m} |\mathbb{E}[Z_{i_1, \dots, i_m} Z_{k_1, \dots, k_m}]|$$

$$\leq \frac{1}{N^2} C \left(\prod_{r=1}^m M_r \right) L^2 (4CL^{1/m} + 1)^m = O\left(\frac{L^3}{N}\right).$$

For the remaining term, we use the following inequality which is similar to Bolthausen (1982) Lemma 1.

Inequality 1. *If X_s is a mixing random field with finite $(2+\delta)^{th}$ moments satisfying (i) $\sum_{m=1}^{\infty} m\alpha_{k,l}(m) < \infty$ for $k+l \leq 4$, (ii) $\alpha_{1,\infty}(m) = o(m^{-2})$, and (iii) For some $\delta > 0$, $\sum_{m=1}^{\infty} m\alpha_{1,1}(m)^{\delta/(2+\delta)} < \infty$, then*

$$|\text{cov}(X_s, X_{s'})| \leq C\alpha_{\infty,\infty}(\text{dist}(s, s'))^{\delta/(2+\delta)} \|X_s\|_{2+\delta} \|X_{s'}\|_{2+\delta}.$$

Applying Inequality 1 to the expectation term in (6.4), we get that

$$|\mathbb{E}[Z_{i_1, \dots, i_m} Z_{k_1, \dots, k_m}]| \leq C\alpha_{\infty,\infty}(kL^{1/m})^{\delta/(2+\delta)} L^{(2m)/m}$$

where we have used that $\|Z_{i_1, \dots, i_m}\|_{2+\delta}^2 \leq CL^2$ which follows under Assumption 8.(iv). There are at most CN points which lie far enough from any given point to fall within the sum in the second term, so we can bound the second term as

$$\begin{aligned} & \frac{1}{N^2} \sum_{i_1}^{M_1} \cdots \sum_{i_m}^{M_m} \sum_{|k_1 - i_1| > 2CL^{1/m} \text{ or } \dots \text{ or } |k_m - i_m| > 2CL^{1/m}} |\mathbb{E}[Z_{i_1, \dots, i_m} Z_{k_1, \dots, k_m}]| \\ & \leq \frac{1}{N^2} \left(\prod_{r=1}^m M_r \right) (CN) \alpha_{\infty,\infty}(kL^{1/m})^{\delta/(2+\delta)} L^{(2m)/m} = C\alpha_{\infty,\infty}(kL^{1/m})^{\delta/(2+\delta)} L^{(2m)/m} \\ & = o(1) \end{aligned}$$

where the last equality follows from the mixing condition in Assumption 8.(iv).²⁹

It now follows immediately that $\left(\frac{1}{N} \sum_{g=1}^G x'_g \varepsilon_g \varepsilon'_g x_g - \tilde{V}_{HAC,L} \right) \xrightarrow{p} 0$.

²⁹Note that the argument here shows that $\tilde{V}_{HAC,L} - \mathbb{E}\tilde{V}_{HAC,L} \xrightarrow{p} 0$. We would then have $\tilde{V}_{HAC,L} \xrightarrow{p} V$ by showing that $V - \mathbb{E}\tilde{V}_{HAC,L} = o(1)$ which would follow from an argument similar to that used to bound the second term in B_N^* .

We now show that $R_1 \xrightarrow{p} 0$. The proof that $R_2 \xrightarrow{p} 0$ is similar and is omitted.

$$\begin{aligned} \text{vec}(R_1) &= \frac{1}{N} \left[\sum_{g=1}^G (x'_g \varepsilon_g \otimes x'_g x_g) \right] \text{vec}(\hat{\beta} - \beta) \\ &= \frac{1}{N^{3/2}} \left[\sum_{g=1}^G (x'_g \varepsilon_g \otimes x'_g x_g) \right] O_p(1) \\ &= \frac{1}{\sqrt{G}} \frac{1}{G} \left[\sum_{g=1}^G (x'_g \varepsilon_g / \sqrt{L} \otimes x'_g x_g / L) \right] \end{aligned}$$

where the second equality follows since the conditions of Assumption 8 give $\hat{\beta} - \beta = O_p(N^{-1/2})$ from Jenish and Prucha (2007) Theorems 1 and 3.

Now consider a typical element of $\frac{1}{G} \left[\sum_{g=1}^G (x'_g \varepsilon_g / \sqrt{L} \otimes x'_g x_g / L) \right]$ given by

$$\tilde{R}_1 = \frac{1}{G} \left[\sum_{g=1}^G \left(\frac{1}{\sqrt{L}} \sum_{s \in \mathcal{G}_g} x_s^* \varepsilon_s \frac{1}{L} \sum_{s \in \mathcal{G}_g} x_s^* x_s^* \right) \right]$$

where the stars refer to an arbitrary element of each vector. We then have

$$\begin{aligned} \tilde{R}_1 &= \frac{1}{G} \left[\sum_{g=1}^G \left(\frac{1}{\sqrt{L}} \sum_{s \in \mathcal{G}_g} x_s^* \varepsilon_s \frac{1}{L} \sum_{s \in \mathcal{G}_g} (x_s^* x_s^* - \mathbb{E}[x_s^* x_s^*]) \right) \right] + \frac{1}{G} \left[\sum_{g=1}^G \left(\frac{1}{\sqrt{L}} \sum_{s \in \mathcal{G}_g} x_s^* \varepsilon_s \frac{1}{L} \sum_{s \in \mathcal{G}_g} \mathbb{E}[x_s^* x_s^*] \right) \right] \\ &\leq \frac{1}{G} \left[\sum_{g=1}^G \left(\frac{1}{\sqrt{L}} \sum_{s \in \mathcal{G}_g} x_s^* \varepsilon_s \frac{1}{L} \sum_{s \in \mathcal{G}_g} (x_s^* x_s^* - \mathbb{E}[x_s^* x_s^*]) \right) \right] + \frac{C}{\sqrt{G}} \frac{1}{\sqrt{N}} \sum x_s^* \varepsilon_s \end{aligned}$$

(6.5)

$$= \frac{1}{G} \left[\sum_{g=1}^G \left(\frac{1}{\sqrt{L}} \sum_{s \in \mathcal{G}_g} x_s^* \varepsilon_s \frac{1}{L} \sum_{s \in \mathcal{G}_g} (x_s^* x_s^* - \mathbb{E}[x_s^* x_s^*]) \right) \right] + \frac{C}{\sqrt{G}} O_p(1)$$

where the first inequality follows under the moment condition in Assumption 8.(iv) and the second equality follows from Jenish and Prucha (2007) Theorem 1 as above.

We then have

$$\mathbb{E} \left| \frac{1}{G} \left[\sum_{g=1}^G \left(\frac{1}{\sqrt{L}} \sum_{s \in \mathcal{G}_g} x_s^* \varepsilon_s \frac{1}{L} \sum_{s \in \mathcal{G}_g} (x_s^* x_s^* - \mathbb{E}[x_s^* x_s^*]) \right) \right] \right|$$

$$\begin{aligned}
&\leq \frac{1}{\sqrt{L}} \frac{1}{G} \sum_{g=1}^G \left[\mathbb{E} \left| \frac{1}{\sqrt{L}} \sum_{s \in \mathcal{G}_g} x_s^* \varepsilon_s \right|^2 \mathbb{E} \left| \frac{1}{\sqrt{L}} \sum_{s \in \mathcal{G}_g} (x_s^* x_s^* - \mathbb{E}[x_s^* x_s^*]) \right|^2 \right]^{1/2} \\
(6.6) \quad &\leq \frac{1}{\sqrt{L}} \frac{1}{G} \sum_{g=1}^G C = \frac{1}{\sqrt{L}}
\end{aligned}$$

where the first inequality follows from the triangle inequality and CS and the second inequality from bounding $\mathbb{E} \left| \frac{1}{\sqrt{L}} \sum_{s \in \mathcal{G}_g} x_s^* \varepsilon_s \right|^2 \leq C$ and $\mathbb{E} \left| \frac{1}{\sqrt{L}} \sum_{s \in \mathcal{G}_g} (x_s^* x_s^* - \mathbb{E}[x_s^* x_s^*]) \right|^2 \leq C$ using standard mixing inequalities, e.g. Jenish and Prucha (2007) Lemma 1 or Bolthausen (1982) Lemma 1.

We illustrate $\mathbb{E} \left| \frac{1}{\sqrt{L}} \sum_{s \in \mathcal{G}_g} x_s^* \varepsilon_s \right|^2 \leq C$ in the following; $\mathbb{E} \left| \frac{1}{\sqrt{L}} \sum_{s \in \mathcal{G}_g} (x_s^* x_s^* - \mathbb{E}[x_s^* x_s^*]) \right|^2 \leq C$ is similar. First note that there are at most $C(m)d$ neighbors within distance d of any given point where $C(m)$ is a constant that depends on the dimension of the index set. Also, from Bolthausen (1982) Lemma 1 and using the moment conditions in Assumption 8.(iv), we have $|\mathbb{E}[x_{s_1}^* \varepsilon_{s_1} x_{s_2}^* \varepsilon_{s_2}]| \leq C\alpha_{1,1}(d)^{\frac{\delta}{2+\delta}}$ where $d = \text{dist}(s_1, s_2)$.

$$\begin{aligned}
\mathbb{E} \left| \frac{1}{\sqrt{L}} \sum_{s \in \mathcal{G}_g} x_s^* \varepsilon_s \right|^2 &= \frac{1}{L} \sum_{s_1 \in \mathcal{G}_g} \sum_{s_2 \in \mathcal{G}_g} \mathbb{E}[x_{s_1}^* \varepsilon_{s_1} x_{s_2}^* \varepsilon_{s_2}] \\
&\leq \frac{1}{L} \sum_{s_1 \in \mathcal{G}_g} \sum_{s_2 \in \mathcal{G}_g} |\mathbb{E}[x_{s_1}^* \varepsilon_{s_1} x_{s_2}^* \varepsilon_{s_2}]| \\
&\leq \frac{1}{L} \sum_{s_1 \in \mathcal{G}_g} \sum_{d=1}^{\infty} C d \alpha_{1,1}(d)^{\frac{\delta}{2+\delta}} \leq C
\end{aligned}$$

where the last inequality follows from the mixing condition in Assumption 8.(iv).

Combining (6.5) and (6.6) gives that $\tilde{R}_1 = O_p(1/\sqrt{L} + 1/\sqrt{G})$ and it follows that $R_1 = O_p(1/\sqrt{N} + 1/G)$. By a similar argument, it also follows that $R_2 = O_p(1/N + 1/G + 1/\sqrt{GN})$. Combining these results with $\left(\frac{1}{N} \sum_{g=1}^G x'_g \varepsilon_g \varepsilon'_g x_g - \tilde{V}_{HAC,L} \right) \xrightarrow{p} 0$ and $\tilde{V}_{HAC,L} \xrightarrow{p} V$ then yields the result. ■

6.3. 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 $g = 1, \dots, G$ 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 $g = 1, \dots, G$ from which it follows that the array $\left(\frac{1}{\sqrt{L}}x'_1 \varepsilon_1, \dots, \frac{1}{\sqrt{L}}x'_G \varepsilon_G\right)' \xrightarrow{d} Z = N(0, W)$ 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 $L \rightarrow \infty$ of

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

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

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

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. Additional Simulation Evidence

This section presents a set of simulation experiments contrasting performance of t -tests using our approach, referred to below as BCH, with that of the Fama-Macbeth t -test of Ibragimov and Müller (2006), hereafter 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, where we discuss the relationship between our fixed- G results and the approach in IM, 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 modest degree of serial correlation ($\rho = \frac{1}{2}$),

$$\begin{aligned}
 Y_t &= b_0 + b_1 X_t + \varepsilon_t \\
 \varepsilon_t &= \rho \varepsilon_{t-1} + u_t \\
 X_t &= \rho X_{t-1} + v_t \\
 \rho &= 1/2 \\
 (u_t, v_t) &\sim N(0, I_2),
 \end{aligned}$$

where v_t is IID $N(0, 1)$ and independent of u_t which is also standard normal. The true parameter values are: $b_0 = 0, b_1 = 1$. Initial conditions are drawn from stationary distributions and point estimates obtained by OLS. This design is in all respects similar to the time series example in Section 4.1.

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:

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

$$\dim(Z) = k$$

$$Z_t = \rho Z_{t-1} + W_t, \quad W_t \sim N \left(0, \frac{1}{k} I_k \right)$$

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 intermittent (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

$$\begin{aligned}
Z_{1t} &= 1(V_t = 0)S_t + V_t \\
V_t &= \begin{cases} 0 & \text{with prob } 96\% \\ +J & \text{with prob } 2\% \\ -J & \text{with prob } 2\% \end{cases} \\
S_t &= \begin{cases} \rho S_{t-1} + w_t & \text{when } V_{t-1} = 0 \\ \rho S_{t-2} + w_t & \text{when } V_{t-1} = \pm J \end{cases}
\end{aligned}$$

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

Results

Appendix Table 1 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.

Appendix Table 2 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 $J = 5, 10, \text{ and } 20$. These values of J result in standard deviations of the sample variance of Z_{1t} across groups of .064, .22, and .90. For the strong instrument setting, $\Pi = 2$, shown in the top panel of the table, BCH suffers from positive 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.

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, 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.
- 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.
- (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.
- (2005): “A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests,” *Econometric Theory*, 21, 1130–1164.
- LEE, L.-F. (2004): “Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Econometric Models,” *Econometrica*, 72, 1899–1926.
- (2007a): “GMM and 2SLS Estimation of Mixed Regressive, Spatial Autoregressive Models,” *Journal of Econometrics*, 137, 489–514.
- (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.

- NEWKEY, W. K., AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics. Volume 4*, ed. by R. F. Engle, and D. L. McFadden. Elsevier: North-Holland.
- NEWKEY, W. K., AND K. D. WEST (1987): "A Simple, Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55(3), 703–708.
- PHILLIPS, P. C. B., Y. SUN, AND S. JIN (2003): "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests," Mimeo.
- 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.
- VOGELSANG, T. J. (2003): "Testing in GMM Models without Truncation," in *Advances in Econometrics Volume 17, Maximum Likelihood Estimation of Misspecified Models: Twenty Years Later*, ed. by T. B. Fomby, and R. C. Hill. Elsevier Science: New York.
- WHITE, H. (2001): *Asymptotic Theory for Econometricians*. San Diego: Academic Press, revised edn.
- WOOLDRIDGE, J. M. (1994): "Estimation and Inference for Dependent Processes," in *Handbook of Econometrics. Volume 4*, ed. by R. F. Engle, and D. L. McFadden. Elsevier: North-Holland.
- (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

	Ref. Dist.	Time Series				Spatial	
		$\rho=0.0$	$\rho=0.5$	$\rho=0.8$	$\gamma=0.0$	$\gamma=0.3$	$\gamma=0.6$
IID	N(0,1)	0.049	0.127	0.341	0.042	0.376	0.538
Heteroskedasticity	N(0,1)	0.056	0.137	0.364	0.044	0.378	0.550
Bartlett-Large H	N(0,1)	0.086	0.114	0.184	0.084	0.114	0.136
Bartlett-Large H	KV	0.023	0.038	0.082			
Bartlett-Med. H	N(0,1)	0.074	0.105	0.191	0.066	0.098	0.116
Bartlett-Med. H	KV	0.044	0.071	0.142			
Bartlett-Small H	N(0,1)	0.063	0.111	0.253	0.044	0.116	0.152
Bartlett-Small H	KV	0.050	0.092	0.227			
CCE-Large L	t(G-1)	0.053	0.059	0.082	0.046	0.032	0.058
CCE-Med. L	t(G-1)	0.056	0.072	0.118	0.054	0.062	0.088
CCE-Small L	t(G-1)	0.058	0.081	0.157	0.040	0.140	0.178

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 4, 8, and 12 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 4, 8, and 16 for HAC and denote numbers of groups (G) of 144, 16, and 4 for CCE.

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

	Ref. Dist.	$\gamma=.8$		$\gamma=.2$	
		$\rho=.8$	$\rho=.4$	$\rho=.8$	$\rho=.4$
IID	N(0,1)	0.683	0.412	0.601	0.336
Heteroskedasticity	N(0,1)	0.685	0.418	0.604	0.337
Cluster:					
State	t(48)	0.253	0.245	0.162	0.144
Month	t(383)	0.489	0.188	0.499	0.199
State/Month	t(48)	0.190	0.117	0.143	0.083
G4 x T3	t(11)	0.103	0.096	0.084	0.081
G4 x T6	t(23)	0.121	0.115	0.108	0.091
G4 x T32	t(127)	0.166	0.113	0.163	0.091
G2 x T3	t(5)	0.083	0.074	0.084	0.063
G2 x T6	t(11)	0.101	0.087	0.099	0.083
G2 x T32	t(63)	0.154	0.103	0.146	0.084
T3	t(2)	0.052	0.059	0.059	0.051
T6	t(5)	0.066	0.051	0.070	0.062
T32	t(31)	0.109	0.065	0.114	0.062
G4	t(3)	0.066	0.063	0.061	0.066
G2	t(1)	0.068	0.050	0.063	0.062
State x T3	t(146)	0.271	0.254	0.180	0.160
State x T6	t(493)	0.284	0.261	0.196	0.175
State x T32	t(1567)	0.344	0.259	0.248	0.174

Note: The table reports rejection rates for 5% level tests from a Monte Carlo simulation experiment with BLS unemployment data regressed on state and month dummies and a randomly generated continuous treatment. All results are based on 1000 simulation replications. The parameters ρ and γ respectively control the strength of the time series and cross-sectional correlation; see text for details. 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 partitioning 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 $T=384$.

Table 3. Simulation Results from Unemployment Data. Discrete Treatment.

	Ref. Dist.	$\gamma=.8$	$\gamma=.4$	$\gamma=.2$	$\gamma=0$
IID	N(0,1)	0.805	0.815	0.806	0.755
Heteroskedasticity	N(0,1)	0.801	0.817	0.802	0.754
Cluster:					
State	t(48)	0.176	0.15	0.111	0.058
Month	t(383)	0.757	0.793	0.807	0.773
State/Month	t(48)	0.171	0.146	0.11	0.059
G4 x T3	t(11)	0.183	0.162	0.164	0.121
G4 x T6	t(23)	0.186	0.17	0.179	0.155
G4 x T32	t(127)	0.393	0.427	0.429	0.402
G2 x T3	t(5)	0.193	0.177	0.156	0.127
G2 x T6	t(11)	0.195	0.165	0.178	0.145
G2 x T32	t(63)	0.405	0.423	0.422	0.405
T3	t(2)	0.112	0.113	0.089	0.106
T6	t(5)	0.133	0.135	0.125	0.132
T32	t(31)	0.365	0.382	0.41	0.403
G4	t(3)	0.082	0.068	0.074	0.059
G2	t(1)	0.071	0.062	0.055	0.05
State x T3	t(146)	0.234	0.234	0.183	0.119
State x T6	t(493)	0.268	0.261	0.208	0.145
State x T32	t(1567)	0.469	0.48	0.449	0.376

Note: The table reports rejection rates for 5% level tests from a Monte Carlo simulation experiment with BLS unemployment data regressed on state and month dummies and a randomly generated binary treatment. All results are based on 1000 simulation replications. The parameter γ controls the strength of the time series and cross-sectional correlation; see text for details. 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 partitioning 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=384.

Figure 1: Power Curve for Test Using CCE with 4 Groups and HAC with Bandwidth 16 and Kiefer-Vogelsang (2005) Reference Distribution for Time Series Simulation with $\rho = 0.8$

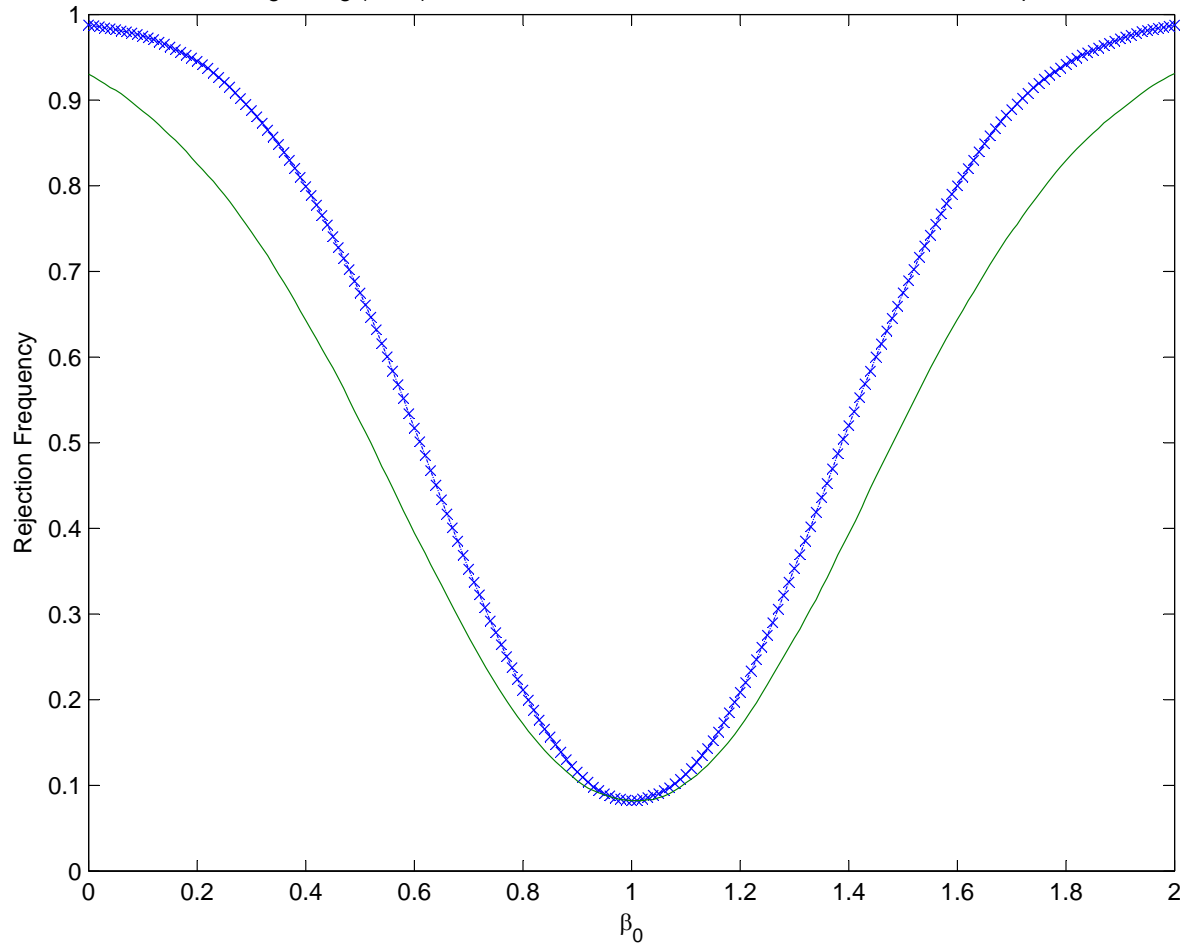
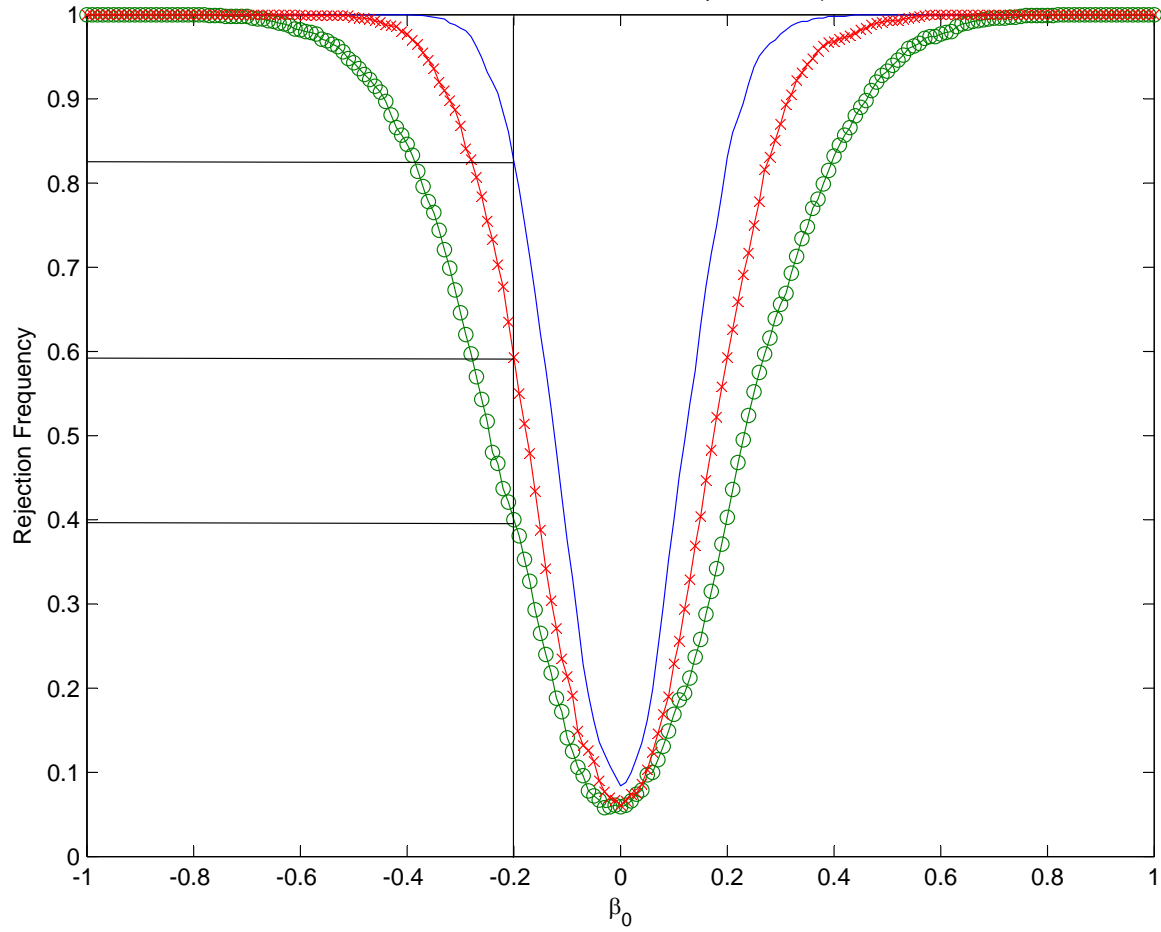


Figure 2: Power Curve for Test Using CCE with G4xT3, G4, and T3 for Unemployment Rate Simulation with Continuous Treatment with $\rho = .8$ and $\gamma = .2$



Appendix Table 1. 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
Benchmark PI=2		BCH	-0.003	0.065	0.061
		IM	-0.013	0.068	0.062
2SLS 3 Instruments 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
PI = 2		BCH	-0.013	0.065	0.067
		IM	-0.049	0.080	0.114
2SLS 6 Instruments 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 $T = 100$. $G = 4$.

Appendix Table 2. 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. Std