

# Improved Spatial Dependence-Robust Inference via Pre-whitening

Timothy G. Conley, Morgan Kelly, and Damian Kozbur \*

November 5, 2025

Preliminary

## 1. Introduction

There are many econometric applications in which observed variables exhibit cross sectional dependence. Failure to account for this dependence when conducting statistical inference may, and typically does, lead to misleading conclusions. Econometric solutions to non-parametrically allow for general forms of cross sectional dependence in either cross section or panel data using spatial models have been around for decades.<sup>1</sup> By non-parametric, we mean methods which allow the flexibility in modelling dependence to be informed by the data. Operationally, non-parametric methods take as input some form of tuning parameter choice which is a function of the data.

Early approaches like [Conley, 1999] use Heteroskedasticity and Autocovariance (HAC) covariance estimators, analogous to those used in time series analysis, that involve a weighted average of sample covariances.<sup>2</sup> To implement these HAC estimators researchers must choose weights that determine which covariances are included in the estimator. In applications where they can be applied, sample splitting/large cluster methods like [Ibragimov and Müller, 2010] and [Bester et al., 2011a] offer potential improvements upon HAC-based inference but they still require a choice of clusters/groups. The most recent methods like [Müller and Watson, 2022] and [Sun and Kim, 2015] with bandwidth calculated using [Lazarus et al., 2018] offer further improvements when applicable, but still require nuisance parameter choices which restrict characteristics of spatial covariance functions.

The associated tuning parameter choice potentially complicates applying any of these methods. Typically, this choice is relatively easy with modest levels of spatial correlation but becomes difficult as the dependence in the data increases. In this paper, we introduce a simple method to make it easier to choose tuning parameters and apply these existing inference methods by reducing the spatial dependence in the covariances that need to be estimated.

---

\*Conley gratefully acknowledges support from the Social Science and Humanities Research Council of Canada. We thank Hans Martinez Torres for outstanding research assistance.

<sup>1</sup>At least since [Conley, 1996] and [Conley, 1999]

<sup>2</sup>See for time series [Bartlett, 1950], [Andrews, 1991].

We illustrate our approach in a linear regression context for ease of exposition. In a linear regression model we include a set of functions of locations as additional regressors. We refer to these extra regressors as spatial basis terms. These spatial basis terms have true coefficients that are zero but they have small-sample correlations that in effect absorb some of the spatial correlation in regression residuals and scores. This reduces the spatial correlations in scores and makes inference easier. We refer to this reduction in spatial dependence as pre-whitening, making scores closer to white noise. Of course, the cost to including spatial basis terms in a regression is that it also reduces the regressor variation that identifies the coefficient(s) of interest. The goal is to trade off a small reduction in identifying variation for an appreciable improvement in spatial dependence inference quality. Our method is not limited to linear regression, it can be easily applied other contexts by simply augmenting conditioning information with spatial basis terms.

We present our method in a context with spatial data indexed on the plane and presume that the researcher has access to a vector of coordinates for each observation. Further we assume that there is a metric that characterizes dependence in the data. We further assume that the data are mixing, which allows us to prove a law of large numbers and central limit theorem. Mixing in this context means that close-by observations can be highly dependent but as distance grows observations approach independence.

There are several ways to generate sensible basis functions which serve to help with spatial pre-whitening. We choose to focus on B-splines as well as higher dimensional basis functions derived either directly or from tensor products of B-splines. One dimensional B-splines are piece-wise polynomials that are nonzero only on a finite range. An order zero B-spline is a step function, and order one B-spline is a piece-wise linear triangle, order two is a piece-wise parabola, etc. B-spline approximations then consist of linear combinations of a collection of these individual B-splines, suitably spread out.

We present a theorem giving conditions over which asymptotic coverage of HAC confidence intervals approaches a nominal value, e.g. 95%. The theorem defines an asymptotic frame over sequences of metric spaces that serve as spatial indexing sets. The spaces/metrics are allowed to be non-Euclidean. We also discuss extensions of our spatial basis approach to large cluster spatial dependence inference methods like those of [Ibragimov and Müller, 2010] and [Bester et al., 2011b]. We also anticipate that our method will be complementary to bootstrap methods, e.g., [Conley et al., 2023], and methods using multiple series with similar covariance structure, e.g., [DellaVigna et al., 2025].

A related contribution to ours is that in [Müller and Watson, 2024], who characterize a class of spatial unit root processes indexed on subsets of a Euclidean plane, demonstrate that classical  $t$  statistics diverge in a suitable sense. They provide a spatial demeaning operation which improves confidence interval coverage distortion problems arising from behavior related to the spurious regression phenomenon.

In Section 2 we present notation and our basic setup, followed by a formal econometric analysis in Section 3. In Section 4 we present a small simulation study that illustrates the inference problem we address and how our approach is a promising solution. In Section 5, we present practical guidance for data-driven choice of spatial pre-whitening bases. We present one method which chooses based upon a

nearest neighbor correlation criteria and an alternative method which is simulation based and similar to the method for choosing cluster numbers in [Cao et al., 2023]. In Section 6, we present an empirical example illustrating the application of our method.

## 2. Data and Estimation

Observed data is a collection of ordered pairs of random variables,  $(Y_i, X_i)$  with  $i$  in an indexing set  $S$ . The  $X_i \in \mathbb{R}^p$  are regressors and  $Y_i \in \mathbb{R}$  are outcome variables. The indexing set  $S$  is observed and has cardinality  $|S| = n$ .  $S$  is also outfitted with a metric or distance measure  $d : S \times S \rightarrow [0, \infty)$ .  $d$  evaluated at  $i$  and  $j$  is denoted  $d_{ij}$ . The definition of  $d$  is extended to subsets  $A, B \subseteq S$  by  $d_{AB} = \inf_{i \in A, j \in B} d_{ij}$ . We will assume below that the data are weakly dependent and that observations  $i$  and  $j$  approach independence as  $d_{ij}$  grows large.

We present our method assuming both  $X_i$  and  $Y_i$  are mean zero for ease of exposition, and we focus on estimation of the linear regression model

$$Y_i = X_i' \beta_0 + \varepsilon_i.$$

The random variables  $\varepsilon_i$  are unobserved and  $\beta_0$  is identified through the usual conditions that  $E[\varepsilon_i X_i] = 0$  and  $E[X_i X_i']$  is full rank. With weakly dependent data, Ordinary Least Squares (OLS) estimates of  $\beta_0$  are consistent.

Given that consistent estimates can be constructed through some estimator  $\hat{\beta}$ , an accompanying problem is constructing a  $1 - \alpha$  level confidence set  $\hat{C}$  for  $\beta_0$ , that satisfies

$$\Pr(\hat{C} \text{ contains } \beta_0) \geq 1 - \alpha - \nu_n$$

where  $\nu_n$  is a remainder which is small in that it can be bounded by a vanishing function of  $n$  for a class of data generating processes which are delimited later.

Failure to account for dependence in the data across  $i$  may lead to substantial distortion of coverage probability (i.e.,  $\Pr(\hat{C} \text{ contains } \beta_0)$  may in practice be far from  $1 - \alpha$ .) Standard methods for constructing  $\hat{C}$  in the context of the linear model with sufficiently strongly mixing properties for observations across  $i$ , is to estimate  $\hat{\beta}$  using least squares estimation, followed by standard error calculation using one of many adjustments for spatial dependence. [Conley, 1999] provides one such example in which a spatial HAC adjustment is used. Subsequent refinements are reviewed above.

We propose a confidence interval procedure which is designed to work together with previously designed spatially robust inferential procedures. Our proposal is to augment the regressors  $X_i$  with additional regressors  $G_i$ , whose construction is described below. That means we will run a regression of the form

$$Y_i = X_i' \beta_0 + G_i' \gamma_0 + \varepsilon_i.$$

The components of  $G_i$  are calculated by evaluating a collection of spatial pre-whitening basis functions at the  $i$ -th location. A spatial pre-whitening basis is a set  $\mathcal{G}$  of functions  $g \in \mathcal{G}$  of the spatial indexing set,  $S$ , each of the form  $g : S \rightarrow [0, 1]$ . The components of  $G_i$  are calculated as  $g(i)$  for all  $g$  in  $\mathcal{G}$ .

The main examples of spatial pre-whitening bases that we discuss below are spatially localized B-splines. When we give practical guidelines for choosing  $\mathcal{G}$  will consider several candidate spline bases, e.g.,  $\mathcal{G}_1, \mathcal{G}_2, \dots$ , in which for instance the knot points might differ.

To construct a confidence interval for a component of  $\beta_0$ , first estimate  $[\hat{\beta}, \hat{\gamma}]$  with an OLS regression of  $Y_i$  on  $[X_i, G_i]$ . Then construct a confidence interval using spatial HAC estimation with bandwidth  $h > 0$  and kernel function  $k$  for the above regression. Let  $\hat{V}_{(k,h)}$  be the corresponding estimate. For a component  $[\beta_0]_j$  of interest of  $\beta_0$ , let  $q_a$  be the  $a$ th quantile of the standard Gaussian random variable (i.e., of  $N(0,1)$ ), and set the total margin of error estimate  $\widehat{\text{m.e.}} = q_{1-\alpha/2}[\hat{V}_{(k,h)}]_{jj}^{-1/2}$ .

$$\hat{C}_j = [\hat{\beta}_j - \widehat{\text{m.e.}}, \hat{\beta}_j + \widehat{\text{m.e.}}].$$

More generally, confidence sets for functionals  $a(\beta_0)$  may be constructed using the delta method the usual way. Confidence ellipsoids covering  $\beta_0$  are also constructed using the usual asymptotic Gaussian approximation. Note that  $\hat{C} = \hat{C}_j \times \mathbb{R}^{p \setminus \{j\}}$  covers all of  $\beta_0$  with the same probability as  $\hat{C}_j$  covers  $[\beta_0]_j$ .

In the sections below, we discuss data-driven choices for pre-whitening basis  $\mathcal{G}$ , kernel  $k$  and bandwidth  $h$ .

### 3. Analysis

We characterize sets of regularity conditions via what we call an asymptotic frame. An asymptotic frame is captured by a pair

$$F = (F_1, F_2)$$

where

$$F_1 = (\ell_{\text{kern}}, \ell_{\text{basis}}) \quad \text{and}$$

$$F_2 = (L_{\text{mom}}, L_{\text{mix}}, L_{\text{cond}}, L_{\text{growth}}, L_{\text{basis}}, L_{\text{kern}})$$

are an ordered pair of vanishing sequences and an ordered tuple of positive constants.

Each of the elements of  $F_2$  is a positive constant dominating measures of regularity of the data and functions used in estimation. They restrict moments, rank, mixing, metric regularity, the kernel, and pre-whitening basis,  $\mathcal{G}$ . Similarly, both elements of  $F_1$  are vanishing sequences of positive real numbers. The parameters in  $F_1$  help characterize the spline basis that we use to augment our regression as well as the size of the HAC bandwidth  $h$  relative to the sample size  $n$ . We demonstrate properties of  $\hat{C}$  defined above relative to a given asymptotic frame.

For any asymptotic frame  $F$ , let  $\mathcal{P}_F$  be a statistical model, which is a collection of random vectors of the form  $(Y_i, X_i)_{i \in S}$ , and each of which satisfies the following conditions.

1. (*Linearity.*)  $Y_i = X_i' \beta_0 + \varepsilon_i$  with  $E[\varepsilon_i | X_j] = 0$  for  $i, j \in S$ ,
2. (*Moments.*)  $|Y_i| + \|X_i\|_2 \leq L_{\text{mom}}$  for  $i \in S$ ,
3. (*Mixing.*) For  $Z_A, Z_B$  depending on  $\{(Y_i, X_i)\}_{i \in A}, \{(Y_i, X_i)\}_{i \in B}$ ,  $A, B \subseteq S$ ,  $Z_B'$  an independent-of- $Z_A$  copy of  $Z_B$  and every function  $v$  taking values in  $[0, 1]$  and depending on two arguments,  $|E[(v(Z_A, Z_B))] - E[v(Z_A, Z_B)']]| \leq 2 \exp(-d_{AB}/L_{\text{mix}})$ .
4. (*Conditioning.*) For all  $R \subseteq S$  nonempty,  $\lambda_{\min}(|R|^{-1} \sum_{i \in R} E[X_i X_i']) \geq 1/L_{\text{cond}}$  and  $\lambda_{\min}(|R|^{-1} E[(\sum_{i \in R} \varepsilon_i X_i) (\sum_{i \in R} \varepsilon_i X_i)']) \geq 1/L_{\text{cond}}$  where  $\lambda_{\min}$  denotes minimum eigenvalue.
5. (*Metric Regularity.*)  $d_{ij} \geq 1$  for  $i, j \in S$  and  $|B_{2r}(i)| \leq L_{\text{growth}} |B_r(i)|$  for  $i \in S, r > 0$  where  $B_r(i)$  is the ball of radius  $r$  about  $i$ .

In addition to assumptions on the data generating process, to each asymptotic frame  $F$ , assign a set of estimation tuning parameters in  $\mathcal{T}_F$  consisting of a kernel function  $k(d)$ , a positive real bandwidth  $h > 0$ , and an association  $S \mapsto \mathcal{G}$  which assigns to every finite metric space  $S$  a collection functions on  $\mathcal{G}$ , which is called a pre-whitening basis, and  $g \in \mathcal{G}$  are of the form  $g : S \rightarrow [0, 1]$ . Let  $\tilde{g}$  be the residual from the linear projection  $g$  on  $\text{span}(\mathcal{G} \setminus \{g\})$ . Estimation parameters in  $\mathcal{T}_F$  satisfy the following conditions.

6. (*Kernel Regularity.*)  $k(0) = 1$ ,  $k(x) = 0$  for  $x \geq 1$  and  $k$  is relative to  $L_{\text{growth}}$  in that  $|1 - K(x)| \leq L_{\text{kern}} x^{L_{\text{growth}}}$ . Also,  $1 \leq (\ell_{\text{kern}})_n h$  and  $h \leq (\ell_{\text{kern}})_n n$ .
7. (*Basis Regularity.*) For  $g \in \mathcal{G}$ ,  $\text{diam}(\text{supp}(g))^2 < (\ell_{\text{basis}})_n h/6$  and  $1 \leq (\ell_{\text{basis}})_n |\text{supp}(g)| \leq (\ell_{\text{basis}})_n L_{\text{basis}} |\{i \in S : |\tilde{g}(i)| > 1/L_{\text{basis}}\}|$ . For  $i \in S$ ,  $|\{g \in \mathcal{G} : i \in \text{supp}(g)\}| \leq L_{\text{basis}}$ . For  $i \in S$  and  $g \in \mathcal{G}$ ,  $|\tilde{g}(i)| \leq L_{\text{basis}}$ .

In the above definition, Condition 1 defines the linear model. Condition 2 states bounds on observable random variables. Condition 3 is a non-degeneracy assumptions on the  $X_i$ . Condition 4 restricts the growth rate of cardinalities of balls within  $S$ . Non-Euclidean metrics are allowed but the growth rate of the number of elements within balls with respect to radius being characterized by bounded doubling as measured by  $L_{\text{growth}}$  is a characteristic that Euclidean spaces do also have.<sup>3</sup> If  $S$  is part of a sequence of cubes in an integer lattice, then  $L_{\text{growth}}$  may be taken to be two raised to a power equal to the dimension of the lattice. Condition 6 restricts attention to  $g$  which have suitably bounded support. This condition models spline-like dictionaries (sets of approximating functions). Finally, Condition 7 imposes standard regularity on the kernel function and bandwidth. A key part of Condition 7 is that  $h$ , the HAC bandwidth, must be longer than  $\text{diam}(\text{supp}(g))$ . The reason for this is that, projecting  $X_i$  data onto spline functions implies nonzero correlations between nearby projection residuals. The HAC bandwidth needs to account for this. Finally, Condition 7 condition that bounds the number of  $g$  supporting any  $i$ . Such a condition holds, for example, in tensor products of B-splines on lattices. For instance, for second order shape preserving B-splines on the interval  $[0, 1]$ , at most 3 spline terms have support over any  $x \in [0, 1]$ .

---

<sup>3</sup>By Assoud's theorem [Assoud, 1977], a regularized version of the metric given by  $S_{**} = (S, d^{1/2})$  admits a bi-Lipshitz embedding into a Euclidean space, where the dimension and bi-Lipshitz constant only depend on the doubling constant, here  $L_{\text{growth}}$ .

There are two main technical hurdles that our analysis needs to handle. First, the pre-whitening regressors are non-stationary (their support is localized), and second, their number may be moderately large (not bounded by an absolute constant, but small relative to  $n$ ). As a result, we design the definition of an asymptotic frame so that the handling of these technical problems must feature in the proof of Theorem 1 below.

**Theorem 1.** For any frame  $F$ , and data generating process in  $\mathcal{P}_F$  and estimation tuning parameters in  $\mathcal{T}_F$  there is a sequence  $\nu$  which depends only on  $F$  which satisfies  $\lim_{n \rightarrow \infty} \nu_n = 0$  and

$$\Pr(\beta_0 \in \widehat{C}) \geq 1 - \alpha - \nu_n.$$

Theorem 1 states that under the statistical model described above, our pre-whitening procedure enjoys coverage of  $1 - \alpha$ , up to a remainder term  $\nu_n$  which vanishes under  $n \rightarrow \infty$ . At the same time, the usual spatial HAC as in [Conley, 1996] also achieves asymptotically  $1 - \alpha$  coverage, again up to a vanishing remainder term. In fact, this can be seen either by referencing the arguments in [Conley, 1996] or by applying Theorem 1 using an empty set of  $g$ .

There are potentially several choices for pre-whitening dictionaries  $\mathcal{G}$ . A good choice of  $\mathcal{G}$  could simultaneously improve coverage probability and reduce the length of the confidence interval. For instance, under the conditions for Theorem 1, HAC estimation without pre-whitening will also have asymptotically correct coverage.

The proof of the Theorem develops properties of  $S$  to derive law of large numbers and central limit theorem -type bounds for spatial data. Central limit theorems have been developed for dependent data, e.g., dating back to [Stein, 1972] or for spatially indexed data more recently in [Jenish and Prucha, 2009]. The bounds we develop require much stronger moment conditions but have the advantage that they depend on  $S$  in an explicit way and only through  $F$ .

The results in Theorem 1 extend to confidence sets constructed using large cluster methods including [Ibragimov and Müller, 2010] and [Bester et al., 2011b]. These methods rely on an approximation that holds for a small (fixed) number of large clusters. These key aspects of this approximation are that within cluster averages are approximately Gaussian and independent of each other. [Cao et al., 2023] demonstrate that a k-medoids clustering algorithm can be used to construct a small set of clusters with large interiors relative to their boundaries that will have these two properties. The mixing properties demonstrated in the proof of Theorem 1 for residuals from projections on our spatial basis terms will hold within-cluster for a small set of large clusters. This, along with moment conditions implies that within-cluster averages are approximately Gaussian and independent of each other. Thus application of [Ibragimov and Müller, 2010] inference is immediate and if the homogeneity restrictions in [Bester et al., 2011b] hold, this method can also be applied.

*Proof of Theorem 1.* Theorem 1 is proven for a scalar  $\beta_0$ ,  $p = 1$ . The proof of the case  $p > 1$  is analogous, noting that information about  $p$  is already implicitly present through the combination of  $(L_{\text{mom}}, L_{\text{cond}})$ .

Let  $\ell = \max((\ell_{\text{kern}})_n, (\ell_{\text{basis}})_n)$ . . Additionally, let  $L = 2 \max(F_2)^8$ .

All log operations are base 2.

**Lemma 1.** For  $T \subseteq S$ ,  $x \geq 1$  and  $\Delta = \{i \in T^2 : d_{i_1 i_2} \leq x\}$ ,  $|\Delta| \leq |T|L^{\log x+2}$ .

*Proof of Lemma 1.* For  $i \in T$  there is the sequence of bounds  $|B_x(i)| \leq L_{\text{growth}}|B_{x/2}(i)| \leq L_{\text{growth}}^2|B_{x/4}(i)| \leq \dots \leq L_{\text{growth}}^{\lceil \log x \rceil + 1}|\{i\}|$ , where  $\lceil x \rceil$  denotes least integer  $\geq x$ . Note  $|\{i\}| = 1$  and  $\lceil \log x \rceil + 1 \leq \log x + 2$ .

Lemma 1 follows with  $|\Delta| = |\cup_{i \in T} (B_x(i) \cap T)| \leq |T|L_{\text{growth}}^{\log x+2} \leq |T|L^{\log x+2}$ .

A law of large numbers is helpful. For  $T \subseteq S$  and  $x \geq 1$  define

$$f_1(T, x, \nu) = 4|T|^{-1}L^{\log x+4} + 8L^2 \exp(-(x - \nu)/L).$$

**Lemma 2.** Let  $W_i$  be random variables at  $i \in T \subseteq S$  with  $\text{var}(W_i) \leq 2L^2$  and  $|\text{corr}(W_i, W_j)| \leq 2 \exp(-(d_{ij} - \nu)/L)$ . Let  $c > 0, x \geq 1$ . Then

$$\Pr \left( |T|^{-1} \left| \sum_{i \in T} W_i - \mathbb{E}[W_i] \right| > c \right) \leq c^{-2} f_1(T, x, \nu).$$

*Proof of Lemma 2.* By Lemma 1,  $|\Delta| \leq |T|L^{\log x+2}$ . Then by partitioning the following sum,  $\mathbb{E}[|T|^{-1}(\sum_{i \in T} W_i - \mathbb{E}[W_i])^2] = |T|^{-2} \mathbb{E}[\sum_{i \in \Delta} (W_{i_1} - \mathbb{E}[W_{i_1}])(W_{i_2} - \mathbb{E}[W_{i_2}]) + \sum_{i \in T^2 \setminus \Delta} (W_{i_1} - \mathbb{E}[W_{i_1}])(W_{i_2} - \mathbb{E}[W_{i_2}])] \leq |T|^{-2}(|\Delta|(2L)^2 + |T|^2(2L)^2 2 \exp(-(x - \nu)/L))$ . Markov's inequality gives the lemma.

Note here that for  $Z_A, Z_B$  and  $Z'_B$  as defined in the mixing condition in the introduction, because  $|\text{corr}| \leq 1$ , it follows that  $|\text{corr}(Z_A, Z_B)| = |\text{corr}(Z_A, Z_B)| - 0 = |\text{corr}(Z_A, Z_B)| - |\text{corr}(Z_A, Z'_B)| \leq \exp(-d_{AB}/L_{\text{mix}})$ .

Next are properties of  $\hat{\xi}, \hat{\eta}$  and  $\hat{\zeta}$ , which are defined as least squares coefficients  $X_i, \varepsilon_i$  and  $Y_i$  on  $G_i$ . Denote  $\tilde{X}_i = X_i - G'_i \hat{\xi}$ ,  $\tilde{\varepsilon}_i = \varepsilon_i - G'_i \hat{\eta}$  and  $\tilde{Y}_i = Y_i - G'_i \hat{\zeta}$ .

**Lemma 3.** For every  $g \in \mathcal{G}$ , there is a set  $K_g$  with  $\text{diam}(K_g) \leq \ell h$  such that  $\hat{\xi}_g, \hat{\eta}$  and  $\hat{\zeta}_g$  depend only on  $X_i$  and  $Y_i$  for  $i \in K_g$ . In addition,  $\tilde{X}_i, \tilde{\varepsilon}_i$  and  $\tilde{Y}_i$  depend only on  $\{(X_i, Y_i)\}_{i \in B_{2\ell h}(i)}$ .

*Proof of Lemma 3.*  $\hat{\xi}_g$  may be found by applying the Frisch Waugh Theorem. Then the least squares solution is  $\hat{\xi}_g = (\sum_{i \in \tilde{G}} \tilde{g}(i)^2)^{-1} \sum_{i \in \tilde{G}} X_i \tilde{g}(i)$ .  $\tilde{g}$  can also be defined using exclusively the  $g' \in \mathcal{G}$  with common points of support with  $g$  given by  $K_g = \text{supp}(g) \cup \bigcup_{g': \text{supp}(g') \cap \text{supp}(g) \neq \emptyset} \text{supp}(g')$  and  $\hat{\xi}_g$  depends only on  $X_i$  for  $i \in K_g$ . As  $\text{supp}(g) \subseteq B_{\ell h/6}(i_g)$  for some  $i_g \in S$ , and as  $\text{diam}(\text{supp}(k)) \leq 2\ell h/6$ , then  $K_g \subseteq B_{\ell h}(i_g)$  and by Lemma 1,  $|K_g| \leq L^{\log \ell h+2}$ . The same holds for  $\hat{\eta}_g$  and  $\hat{\zeta}_g$ .

**Lemma 4** For random variables  $Z_A, Z_B$  which depend only on  $\{(\tilde{Y}_i, \tilde{X}_i)\}_{i \in A}$  and  $\{(\tilde{Y}_i, \tilde{X}_i)\}_{i \in B}$ , for  $A, B \subseteq S$ ,  $Z'_B$  an independent-of- $Z_A$  copy of  $Z_B$  and  $v \in [0, 1]$  depending on two arguments,

$$|\mathbb{E}[v(Z_A, Z_B)] - \mathbb{E}[v(Z_A, Z'_B)]| \leq 2 \exp(-(d_{AB} - 4\ell h)/L_{\text{mix}}).$$

*Proof of Lemma 4* Events depending on  $Z_A$  can be defined using  $\{(Y_i, X_i)\}_{i \in A^{2\ell h}}$  where  $A^{2\ell h}$  is the  $2\ell h$  enlargement of  $A$  given by  $\{i \in S : d_{iA} \leq 2\ell h\}$ . The same is true for  $Z_B$ . Then apply the mixing condition from the body of the paper and note that  $d_{A^{2\ell h} B^{2\ell h}} \geq d_{AB} - 4\ell h$ .

**Lemma 5.** Let  $x, y \geq 1$ . Define subsets of  $S^4$ :

$$\begin{aligned} A &= \{i \in S^4 : d_{i_1 i_2} \leq y \text{ and } d_{i_3, i_4} \leq y\}, \\ C_1 &= \{i \in A : \text{diam}(\{i_1, i_2, i_3, i_4\}) \leq 3x\}, \\ C_2 &= \{i \in A \setminus C_1 : d_{\pi i_1, \{\pi i_2, \pi i_3, \pi i_4\}} \geq x \text{ for some permutation } \pi\}, \\ C_3 &= \{i \in A \setminus (C_1 \cup C_2) : d_{\{\pi i_1, \pi i_2\}, \{\pi i_3, \pi i_4\}} \geq x \text{ for some permutation } \pi\}. \end{aligned}$$

Then  $C_1 \cup C_2 \cup C_3 = A$  and

$$|C_1| \leq nL^{3 \log 3x+6} \quad \text{and} \quad |C_2| + |C_3| \leq |A| \leq n^2 L^{2 \log y+4}.$$

*Proof of Lemma 5.* To show the first statement suppose  $i \in A, i \notin C_1 \cup C_2$ . There must be  $\pi$  such that  $d_{\pi i_1 \pi i_3} > 3x$ . As  $i \notin C_2$ , both  $B_r(\pi i_1)$  and  $B_r(\pi i_3)$  must each contain a remaining component of  $i$ , which may be taken  $\pi i_2$  and  $\pi i_4$  respectively. By triangle inequality  $d_{\pi i_2 \pi i_4} > x$  as well as  $d_{\{\pi i_1, \pi i_2\}, \{\pi i_3, \pi i_4\}} > r$ . So  $i \in C_3$ . Next bound the cardinalities of  $A, C_1$ . Let  $A^{1/2} = \{i \in S^2 : i_2 \in B_y(i_1)\}$ . Then  $|A^{1/2}| \leq n \max_{i \in S} |B_y(i)|$ . As in Lemma 1,  $|B_y(i)| \leq L^{\log y+2}$ . Then  $A = A^{1/2} \times A^{1/2}$  gives  $|A| \leq |A^{1/2}|^2$ . Similarly,  $|C_1|$  is bounded analogously. Finally, by inclusion,  $|C_2| + |C_3| \leq |A|$ , and the lemma is proven.

Let  $f_2(R, x) = L \exp(-(x/3 - 2\ell h)/L) + 4!|R|^{-2} L^{2 \log x+5}$ .

**Lemma 6** Let  $z_i$  be mean 0 random variables with  $E[z_i^4] \leq L$ , let  $R \subseteq S$  and let  $W = |R|^{-1} \sum_{i \in R} z_i$ . Let  $x \geq 1$ . Then  $E[W^4] \leq f_2(R, x)$ .

*Proof of Lemma 6.* Let  $x \geq 1$  and let  $A^\circ = \{i \in R^4 : \text{no permutation of } i \text{ is in } A\}$  where  $A$  is the set defined in Lemma 5 using  $y = x$ . If  $i \in A^\circ$  then there is a permutation of  $i$  such that  $d_{\pi i_1, \{\pi i_2, \pi i_3, \pi i_4\}} \geq x/3$ . To see this, note either  $d_{i_1 i_2} \geq x$  or  $d_{i_3 i_4} \geq x$ . If the first, then by also either  $d_{i_1 i_3} \geq x$  or  $d_{i_2 i_4} \geq x$ , either  $d_{i_1 \{i_2, i_3\}} \geq x$  or  $d_{i_2, \{i_1, i_4\}} \geq x$ . In the first of these cases, either  $d_{i_1 i_4} \geq x/3$ , in which the desired permutation is the identity, or  $d_{i_1 i_4} \leq x/3$  and so  $d_{i_2 i_3} \geq x$ . By triangle inequality, one of  $i_2$  or  $i_3$  must have distance  $\geq x/3$  from the remaining elements of  $\{i_1, i_2, i_3, i_4\}$ . Then  $\max_{i \in A^\circ} E[z_{i_1} z_{i_2} z_{i_3} z_{i_4}] \leq L \exp(-(x/3 - \ell h)/L)$  while  $\max_{i \in R^4} E[z_{i_1} z_{i_2} z_{i_3} z_{i_4}] \leq L$  by iterated application of Cauchy-Schwarz inequality. Then  $E[W^4] = |R^{-4}|(\sum_{i \in A^\circ} + \sum_{i \notin A^\circ}) E[z_{i_1} z_{i_2} z_{i_3} z_{i_4}] \leq |R^{-4}| |A^\circ| L \exp(-(x/3 - \ell h)/L) + |R^{-4}| 4! |R|^2 L^{2 \log x+4}$ . Using Lemma 5 to bound the cardinality of the complement of  $A^\circ$ . Simplifying and applying  $|R^{-4}| |A^\circ| \leq 1$  gives the proof.

**Lemma 7.** For  $g \in \mathcal{G}$ ,  $E[\widehat{\xi}_g^4] \leq L^4 f_2(\text{supp}(\tilde{g}), x)$ . Additionally,  $E[(G'_i \widehat{\xi})^4] \leq L^7 \max_{g \in \mathcal{G}} f_2(\text{supp}(\tilde{g}), x)$ . Finally,  $E[\tilde{X}_i^4] \leq 4L^7 (1 + \max_{g \in \mathcal{G}} f_2(\text{supp}(\tilde{g}), x))$ .

*Proof of Lemma 7.* Apply Lemma 6 to  $\widehat{\xi}_g$  making sure to account for the least squares solution denominator and to lower bound using the dictionary regularity condition. Also,  $E[(G'_i \widehat{\xi})^4] = E[(\sum_{g: \text{supp}(g) \ni i} g(i) \widehat{\xi}_g)^4] \leq |\{g : \text{supp}(g) \ni i\}|^2 \sum_{g: \text{supp}(g) \ni i} g(i)^4 E[\widehat{\xi}_g^4] \leq L^7 \max_{g \in \mathcal{G}} f_2(\text{supp}(\tilde{g}), x)$ . Finally,  $\tilde{X}_i = X_i - G'_i \widehat{\xi}$  so  $E[\tilde{X}_i^4] \leq 4(L^4 + \max_{g \in \mathcal{G}} L^7 f_2(\text{supp}(\tilde{g}), x))$ .

Next is a law of large numbers for  $\tilde{X}_i^2$ .

**Lemma 8.** For  $x \geq 1$  and  $c > 0$ ,

$$\Pr(|n^{-1} \sum_{i \in S} \tilde{X}_i^2 - \mathbb{E}[X_i^2]| \geq c + L^7 \max_{g \in \mathcal{G}} f_2(\text{supp}(\tilde{g}), x)^{1/2}) \leq 8c^{-2} f_1(S, x, 4\ell h).$$

*Proof of Lemma 8.* By Lemma 2, for  $x \geq 1$ ,  $\Pr(|n^{-1} \sum_{i \in S} X_i^2 - \mathbb{E}[n^{-1} \sum_{i \in S} X_i^2]| \geq c/2) \leq 4c^{-2} f_1(S, x, 0)$ . By least squares optimality,  $\sum_{i \in S} X_i \tilde{X}_i = 0$ . Lemma 2 also applies to produce the additional bound  $\Pr(n^{-1} \sum_{i \in S} (G_i \hat{\xi})^2 - (L^7 \max_{g \in \mathcal{G}} f_2(\text{supp}(\tilde{g}), x))^{1/2} \geq c/2) \leq 4c^{-2} f_1(S, x, 4\ell h)$ . Combining and simplifying gives the lemma.

Also needed is a central limit theorem for  $\tilde{X}_i \varepsilon_i$ , which is next. Let  $S_{**} \subseteq E$  be the image of a bi-Lipschitz Euclidean embedding  $\iota$  to a Euclidean with bi-Lipschitz constant and dimension depending only on  $L$ , which exists by Assoud's theorem; see description in previous section. The proof of the following lemma constructs a function

$$f_3(S) \text{ with } \lim_{|S| \rightarrow \infty} f_3(S)$$

which depends only on  $S$  and  $F$  and depends on  $S$  only through  $|S|$ .

**Lemma 9.** Let  $\Xi = \sum_{i \in S} \tilde{X}_i \varepsilon_i$ . Let  $\sigma^2 = \text{var}(\Xi)$ . Let  $t \in \mathbb{R}$ . Then

$$\Pr(\sigma^{-1} \Xi \leq t) - \Pr(N(0, 1) \leq t) \leq f_3(S).$$

*Proof of Lemma 9.* Let  $0 < a < 1$  and  $Q_0 = [0, m]^{\dim(\iota S)}$  and  $U_0 = Q_0 \setminus [0, (1-a)m]^{\dim(\iota S)}$ . Let  $U = U_0 + (m\mathbb{Z})^{\dim(\iota S)}$ . Then by the pigeonhole principle there is  $w \in \{0, 1, \dots, m\}^{\dim(\iota S)}$  such that  $|(w + U) \cap \iota S_{**}| \leq n^{1/\bar{L}} \bar{L}$  where  $\bar{L}$  may depend on  $L$  as well as  $\dim(\iota S)$  and the bi-Lipschitz constant of  $\iota$  and  $a$ . Then there is a collection  $\mathcal{R}$  of  $|\mathcal{R}| = m$  disjoint subsets such that for  $R, R' \in \mathcal{R}$ ,  $d_{R, R'} \geq am$  and  $|S \setminus \cup_{R \in \mathcal{R}} R| \leq \bar{L} n^{1/\bar{L}}$  and for which  $m \geq n^{1/\bar{L}} / \bar{L}$ . By taking unions of  $R, R' \in \mathcal{R}$  if necessary, all  $R$  may be taken to have  $|R|/L \leq |R'| \leq L|R|$ . For  $R \in \mathcal{R}$  let  $W_R = \sum_{i \in R} \tilde{X}_i \varepsilon_i$ . Equate  $\Xi = \sum_{R \in \mathcal{R}} W_R + r$  for a remainder  $r$ . Let  $W'_R$  be independent copies of  $W_R$ . Order  $R \in \mathcal{R}$  arbitrarily with  $R_1, \dots, R_m$ . Then let  $\Xi_0 = \Xi - r$  and  $\Xi_l = \Xi_{l-1} - W_{R_l} + W'_{R_l}$ . Then  $\Xi_m$ , by the Berry-Esseen central limit theorem, satisfies  $\Pr(\sigma_m^{-1} \Xi_m \leq t) - \Pr(N(0, 1) \leq t) \leq m^{-1/2} \max_{R \in \mathcal{R}} \mathbb{E}[|W_R|^3] \max_{R \in \mathcal{R}} \mathbb{E}[W_R^2]^{-3/2}$  where  $\sigma_m$  is the variance of  $\Xi_m$ . To bound 3rd moments of sums of  $z_i$ , refer to Lemma 4 above. Then  $\mathbb{E}[|W_R|^3] \leq \mathbb{E}[|W_B|^4]^{3/4} \leq \max_{R \in \mathcal{R}, x \geq 1} (|R|^4 L^8 \exp(-(x/3 - 2\ell h)/L) + 4!|R|^2 L^{2 \log x + 4} L^8)^{3/4}$ . Conversely,  $\mathbb{E}[W_R^2]$  is lowerbounded by  $1/L$  using the conditioning regularity conditions. Finally,  $|\Pr(\Xi_l \leq t) - \Pr(\Xi_{l-1} \leq t)| \leq 2 \exp(am - 4\ell h)/L$ . Summing over  $l$  and optimizing over  $a, x$  and accounting for  $r$  provides for the existence of  $f_3(S)$ .

$$\text{Let } f_4(x) = n^{-1} L^{3 \log 3h+6} L^8 + L^{2 \log x + 4} 3L^9 \exp(-(x - 4\ell h)/L).$$

**Lemma 10.** For any  $x \geq 1$  and  $c > 0$ ,  $\Pr(|\Omega_0^K - \mathbb{E}[\Omega_0^K]| \geq c) \leq c^{-2} f_4(x)$ .

*Proof of Lemma 10.* Define  $A, C_1, C_2, C_3$  as in Lemma 5 and specialize to  $y = h$ . Let  $z_i = \varepsilon_i \tilde{X}_i$ . Then  $\mathbb{E}[(\Omega_0^K - \mathbb{E}[\Omega_0^K])^2]$  expands to

$$\begin{aligned} & \mathbb{E}\left[\frac{1}{n^2} \sum_{i \in S^4} K_{i_1 i_2} K_{i_3 i_4} (z_{i_1} z_{i_2} - \mathbb{E}[z_{i_1} z_{i_2}])(z_{i_3} z_{i_4} - \mathbb{E}[z_{i_3} z_{i_4}])\right] \\ &= \frac{1}{n^2} \sum_{i \in A} K_{i_1 i_2} K_{i_3 i_4} \left( \mathbb{E}[z_{i_1} z_{i_2} z_{i_3} z_{i_4}] - \mathbb{E}[z_{i_1} z_{i_2}] \mathbb{E}[z_{i_3} z_{i_4}] \right). \end{aligned}$$

Let  $M_j = \max_{i \in C_j} K_{i_1 i_2} K_{i_3 i_4} |\mathbb{E}[z_{i_1} z_{i_2} z_{i_3} z_{i_4}] - \mathbb{E}[z_{\pi i_1} z_{\pi i_2}] \mathbb{E}[z_{\pi i_3} z_{\pi i_4}]|$ ,  $j \leq 3$ . Due to the fact that  $2 \max_{i \in C_1} K_{i_1 i_2} K_{i_3 i_4} \mathbb{E}[|z_{i_1} z_{i_2} z_{i_3} z_{i_4}|] \leq L$ ,  $M_1 \leq L$ . Next define a decomposition  $M_2 \leq M_{2a} + M_{2b}$  with first term  $M_{2a} = \max_{i \in C_2} K_{i_1 i_2} K_{i_3 i_4} |\mathbb{E}[z_{i_1} z_{i_2} z_{i_3} z_{i_4}] - \mathbb{E}[z_{\pi i_1}] \mathbb{E}[z_{\pi i_2} z_{\pi i_3} z_{\pi i_4}]|$ , and with second term  $M_{2b} = \max_{i \in C_2} K_{i_1 i_2} K_{i_3 i_4} |\mathbb{E}[z_{\pi i_1}] \mathbb{E}[z_{\pi i_2} z_{\pi i_3} z_{\pi i_4}] - \mathbb{E}[z_{i_1} z_{i_2}] \mathbb{E}[z_{i_3} z_{i_4}]|$ . As  $d_{B_D(i_1), B_D(i_2) \cup B_D(i_3) \cup B_D(i_4)} \geq r$ , applying the mixing from Lemma 6 gives  $M_{2a} \leq L \exp(-(x - 4\ell h)/L)$ . In addition,  $\mathbb{E}[z_{\pi i_1}] = 0$  and either the bound  $|\mathbb{E}[z_{i_1} z_{i_2}]| = |\mathbb{E}[z_{i_1}] \mathbb{E}[z_{i_2}]| \leq L^4 \cdot L \exp(-(x - 4\ell h)/L)$  holds or the same bound for  $(i_3, i_4)$  holds. Together, these give  $M_{2b} \leq 2L \exp(-(r - 4h\ell)/L)$ . For  $M_3$ , if  $\pi \in \langle \{(1\ 2), (3\ 4)\} \rangle$ , then  $|\mathbb{E}[z_{i_1} z_{i_2} z_{i_3} z_{i_4}] - \mathbb{E}[z_{i_1} z_{i_2}] \mathbb{E}[z_{i_3} z_{i_4}]| = |\mathbb{E}[z_{i_1} z_{i_2} z_{i_3} z_{i_4}] - \mathbb{E}[z_{\pi i_1} z_{\pi i_2}] \mathbb{E}[z_{\pi i_3} z_{\pi i_4}]| \leq L \exp(-(x - 4\ell h)/L)$ . If not, then either  $d_{i_1 i_2} \geq x \geq h$  or  $d_{i_3 i_4} \geq x \geq h$  and therefore  $K_{i_1 i_2} = 0$  or  $K_{i_3 i_4} = 0$ . Then  $M_3 \leq L \exp(-(x - 4\ell h)/L)$ .

From the bounds on  $M_1, M_2, M_3, |A|, |C_1|$ , and that  $|C_2|, |C_3| \leq |A|$ ,

$$\begin{aligned} & \mathbb{E}[(\bar{\Omega}_0^K - \mathbb{E}[\bar{\Omega}_0^K])^2] \leq \frac{1}{n^2} |C_1| M_1 + \frac{1}{n^2} |A| M_2 + \frac{1}{n^2} |A| M_3 \\ & \leq \frac{1}{n^2} (nL^{3 \log 3x+6} L^8 + n^2 L^{2 \log h+4} 3L^9 \exp(-(x - 4\ell h)/L)) \end{aligned}$$

Using Markov's inequality and simplifying gives the lemma.

Next let  $\delta_\beta = \beta_0 - \hat{\beta}$ . Denote  $\mathbb{E}_S = n^{-1} \sum_{i_1 \in S}$  and  $\mathbb{E}_S^K = \sum_{i_2 \in S} K_{i_1 i_2}$ . Then

$$\begin{aligned} \hat{\Omega} - \Omega_0^K &= \mathbb{E}_S \mathbb{E}_S^K \bar{X}_{i_1} (\varepsilon_{i_1} + X_{i_1} \delta_\beta - G_{i_1} \hat{\gamma}) \bar{X}_{i_2} (\varepsilon_{i_2} + X_{i_2} \delta_\beta - G_{i_2} \hat{\gamma}) \\ &\quad - \mathbb{E}_S \mathbb{E}_S^K \bar{X}_{i_1} \varepsilon_{i_1} \bar{X}_{i_2} \varepsilon_{i_2}. \end{aligned}$$

For a parameter  $u \in \mathbb{R}$  define

$$\begin{aligned} \delta_1 &= \mathbb{E}_S \mathbb{E}_S^K \tilde{X}_{i_1} X_{i_1} \tilde{X}_{i_2} X_{i_2} u^2, \quad \delta_2 = -2 \mathbb{E}_S \mathbb{E}_S^K \tilde{X}_{i_1} G_{i_1} \hat{\gamma} \tilde{X}_{i_2} X_{i_2} u, \\ \delta_3 &= \mathbb{E}_S \mathbb{E}_S^K \tilde{X}_{i_1} G'_{i_1} \hat{\gamma} \tilde{X}_{i_2} G'_{i_2} \hat{\gamma}, \quad \delta_4 = 2 \mathbb{E}_S \mathbb{E}_S^K \tilde{X}_{i_1} \varepsilon_{i_1} \tilde{X}_{i_2} X_{i_2} u, \\ \delta_5 &= 2 \mathbb{E}_S \mathbb{E}_S^K \tilde{X}_{i_1} \varepsilon_{i_1} \tilde{X}_{i_2} G'_{i_2} \hat{\gamma}. \end{aligned}$$

Under the special case  $u = \delta_\beta$ , the decomposition  $\hat{\Omega} - \Omega_0^K = \delta_1 + \dots + \delta_5$  holds.

Let

$$\begin{aligned} f_5(x) &= 4L^6 (1 + L \max_{g \in \mathcal{G}} f_2(\text{supp}(\tilde{g}), x) L^{\log h+2}) f_1(S, x, 4\ell h + 2h) \\ &\quad + n^3 L^3 2 \exp(-(x - 4\ell h)/L) + L^5 L^{\log x+2} \max_{g \in \mathcal{G}} f_2(\text{supp}(\tilde{g}), x). \end{aligned}$$

**Lemma 11.** For  $j = 1, \dots, 5$ ,  $x \geq 1$  and  $|u| \leq 1$ ,

$$\Pr(\delta_j \geq c \cap \delta_\beta^2 \leq u^2) \leq c^{-2} u f_5(x) + \Pr(\delta_\beta^2 \leq u^2).$$

*Proof of Lemma 11.* Using the 4th moment bound of Lemma 3 with Lemma 2,

$$\Pr(\delta_j \geq c) \leq uc^{-2}4L^6(1 + L \max_{g \in \mathcal{G}} f_2(\text{supp}(\tilde{g}), x)L^{\log h+2})f_1(S, x, 4\ell h + 2h).$$

For the term  $\delta_5$ , note that  $\hat{\gamma}_g = (\sum_{i \in K_g} \tilde{g}(i)^2)^{-1} \sum_{i \in K_g} \tilde{g}(i)Y_i$ . Let  $D_g$  be the denominator and define  $W_{i_1} = \sum_{i_2 \in B_h(i_1)} K_{i_1 i_2} \tilde{X}_{i_1} \varepsilon_{i_1} \tilde{X}_{i_2} G'_{i_2} \hat{\gamma}$  so  $\delta_5 = n^{-1} \sum_{i_1 \in S} W_{i_1}$  and

$$\begin{aligned} \mathbb{E}[W_{i_1}] &= \mathbb{E} \left[ \sum_{i_2 \in B_h(i_1)} K_{i_1 i_2} \tilde{X}_{i_1} \varepsilon_{i_1} \tilde{X}_{i_2} \sum_{g \in \mathcal{G}} g(i_2) \hat{\gamma}_g \right] \\ &= \mathbb{E} \left[ \sum_{i_2 \in B_h(i_1)} K_{i_1 i_2} \tilde{X}_{i_1} \varepsilon_{i_1} \tilde{X}_{i_2} \sum_{g \in \mathcal{G}} g(i_2) D_g^{-1} \left( \sum_{i \in B_x(i_1)} \tilde{g}(i) Y_i + \sum_{i \in K_g \setminus B_x(i_1)} \tilde{g}(i) Y_i \right) \right]. \end{aligned}$$

For  $i \in K_g \setminus B_x(i)$ , note that the above expectation is  $\leq L^2 2 \exp(-(x-4\ell h)/L) \times (|\{g : g(i_2) \neq 0\}| \times |K_g| \times |B_h(i_1)|)$  while the contribution of the  $i \in B_x(i)$  terms is limited to  $|\{g : g(i_2) \neq 0\}| \times |B_x(i)| \times L^4/D_g$  adding to a total bound of

$$\leq n^3 L^3 2 \exp(-(x-4\ell h)/L) + L^5 L^{\log x+2} \max_{g \in \mathcal{G}} f_2(\text{supp}(\tilde{g}), x).$$

For  $\delta_3$ , note that  $\hat{\gamma} = \hat{\xi} \delta_\beta + \hat{\eta}$ . Decompose

$$\delta_3 = \mathbb{E}_S \mathbb{E}_S^K \tilde{X}_{i_1} G'_{i_1} \hat{\eta} \tilde{X}_{i_2} G'_{i_2} \hat{\gamma} + \mathbb{E}_S \mathbb{E}_S^K \tilde{X}_{i_1} G'_{i_1} \hat{\xi} \tilde{X}_{i_2} G'_{i_2} \hat{\xi} \delta_\beta.$$

Formally replacing  $u$  for  $\delta_\beta$  in the right hand term allows proceeding exactly as for  $\delta_2$  to calculate a bound. The left hand term is bounded exactly like  $\delta_3$ .

The Lemma holds after simplifying.

$$\text{Let } f_6(x) = L^{\log x + xh^{-1}L+4} f_3(S, x) + n^2 L \exp(-(x-4\ell h)/L).$$

**Lemma 12.**  $|\mathbb{E}[\tilde{\Omega}_0] - \mathbb{E}[\tilde{\Omega}_0^K]| \leq f_6(x)$ .

*Proof of Lemma 12.* Use  $T = S$  and  $\Delta$  defined with  $x \geq 1$  as in Lemma 1. Then  $|\Delta| \leq nL_{\text{growth}}^{\log x+2}$  and for  $i \in \Delta$ , using the smoothness assumption,  $|1 - K_{i_1 i_2}| \leq L_{\text{kern}}(xh^{-1})^{L_{\text{growth}}}$ . Note that  $|\Delta| \times |1 - K_{i_1 i_2}| \leq nL_{\text{kern}}L_{\text{growth}}^2 L_{\text{growth}}^{\log x + xh^{-1} \log L_{\text{growth}}}$ .

$$\begin{aligned} \mathbb{E}[\Omega_0^K] - \mathbb{E}[\Omega_0] &= n^{-1} \sum_{i \in S^2} \mathbb{E}[(K_{i_1 i_2} - 1) \tilde{X}_{i_1} \varepsilon_{i_1} \tilde{X}_{i_2} \varepsilon_{i_2}] \\ &= \sum_{R \in \{\Delta, S^2 \setminus \Delta\}} n^{-1} \sum_{i \in R} \mathbb{E}[(K_{i_1 i_2} - 1) \tilde{X}_{i_1} \varepsilon_{i_1} \tilde{X}_{i_2} \varepsilon_{i_2}]. \end{aligned}$$

Each sum above is bounded as in the previous lemmas.

Theorem 1 now follows by choosing  $x$  to be an appropriately intermediate sequence depending on  $F$  and combining the probability bounds of the above lemmas. *QED.*

## 4. Simulation Study

This section provides simulation results that illustrate the nature of our inference problem and how our proposed method will improve inference.

Our simulations use a set of  $n = 500$  uniformly distributed locations on a unit square for location data. These locations are drawn once and used for all subsequent simulations. We consider a regression of  $Y_i$  on  $X_i$  where both processes are one dimensional (i.e.,  $X_i, Y_i \in \mathbb{R}$ ), and have the joint same distributions, and are independent of each other. Both variables have the same multivariate Gaussian distribution that can be viewed as a sum of idiosyncratic noise with a spatially correlated component. The DGP is mean zero with a covariance matrix that is a linear combination of a scaled identity matrix and a non-diagonal matrix  $\Sigma$ . So we generate variables  $X_i$  and  $Y_i$  with:

$$X \sim \text{MVN}(0, [(1 - \rho)I + \rho\Sigma]) \quad \text{and} \quad Y \sim \text{MVN}(0, [(1 - \rho)I + \rho\Sigma])$$

Where  $X$  has components  $X_i$  and  $Y$  has components  $Y_i$ .  $\Sigma$  has variances of one and off-diagonal elements  $(i, j)$  given by  $\exp(-d_{ij}^{\text{Euc}}/\theta)$  with  $d_{ij}^{\text{Euc}}$  being the Euclidean distance between locations  $i$  and  $j$ . Again,  $Y_i$  have the same DGP as  $X_i$  and they are independent of each other.

We present results where  $\Sigma$  has parameter  $\theta = \sqrt{2}/10$ . To better understand the level of spatial correlation implied by this value of  $\theta$ , consider the implied ratio of the variance of the sample mean of the elements of a  $N(0, \Sigma)$  vector relative to the analog for an  $N(0, I)$  vector. A  $\theta = \sqrt{2}/10$  implies a sample mean variance that is approximately 45 times greater than if the DGP were  $N(0, I)$ . If the same number of observations were generated from a discrete time series AR1 model, this level of dependence would correspond to an AR1 with slope of approximately .96. Thus, varying the parameter  $\rho$  from zero to one results in a wide variety of dependence levels for  $X_i$  and  $Y_i$ . Furthermore, this type of DGP presents a challenge for HAC estimators even with smaller levels of  $\rho$  since it displays non-trivial correlations for relatively large (compared to our unit square sample region) distances, even when the implied variance of the mean is moderate. To capture enough terms to do well in terms of bias, kernel bandwidths/cutoffs need to be large enough that they have enough noise to potentially undermine the quality of distribution approximations which do not account for noise in variance estimators (and hence do not account for noise in the denominator of t-statistics).

Entries in Table 1 are rejection frequencies for t-tests under the true null hypothesis of zero slope in a regression of  $Y_i$  on  $X_i$ . The first panel presents results with no  $G_i$  terms and different bandwidths using a Gaussian kernel, with variance  $\sigma^2 I$ .<sup>4</sup> The bandwidth is described by headings .05, .10, .15 which give the value of  $2\sigma$  for each kernel. The second panel uses the same HAC estimator but adds an  $8 \times 8$  tensor product of triangular B-splines serving as  $G_i$  to the regression.<sup>5</sup>

<sup>4</sup>For our simulations, we avoid the use of easier-to-interpret uniform kernels since (as is well known) they can yield negative variance estimates and this happens frequently enough to be an issue.

<sup>5</sup>In each coordinate dimension the interior splines are spaced to be shape preserving and a ‘half-triangle’ is used at each edge of the coordinates’ support, see Figure A.1. The tensor product is then formed as all cross-products of these splines in each dimension.

Rows in Table 1 present differing values of  $\rho$ , starting from  $\rho = 0$  when both  $X_i$  and  $Y_i$  are white noise. Subsequent rows present alternative values of  $\rho$ . To illustrate the amount of correlation in both  $X_i$  and  $Y_i$  as  $\rho$  increases, the second column labeled ‘corr’ reports the correlation between pairs of observations at a distance of .10. It is important to note that spatial correlations that would be small in a familiar time series setting can be very substantial in a spatial setting where there are many neighbors at even small distances. Small pairwise correlations can add up to very substantial variation in sample means. As mentioned above, as  $\rho$  approaches one the variance of sample means is similar to its analog for a highly serially correlated AR1 process.

The ‘No Splines’ panel illustrates the HAC difficulties that concern us. Appreciable size distortions are apparent for  $\rho$  values of .2 and above. Size distortions become very severe as  $\rho$  approaches one. Increasing kernel bandwidth/cutoff can help improve size distortions but this alone cannot eliminate distortions because increasing cutoffs while improving bias comes at a cost of increasing noise in variance estimates undermining the quality of the typical spatial HAC [Conley, 1999] variance approximation used here.

The ‘Triangle Splines’ panel presents t-test results for regressions that have been augmented with an 8 by 8 tensor product of the triangle (piece-wise linear) B-splines illustrated in Figure 1. Addition of these B-spline terms can be seen to dramatically improve rejection frequencies, even for the higher values of  $\rho$  that generate data with extremely high levels of spatial correlation. This illustrates the potential for our method to drastically improve the size performance of these HAC methods. The sensitivity of rejection frequencies to bandwidth choice is also greatly reduced. With our method, HAC can work better and be easier to implement.

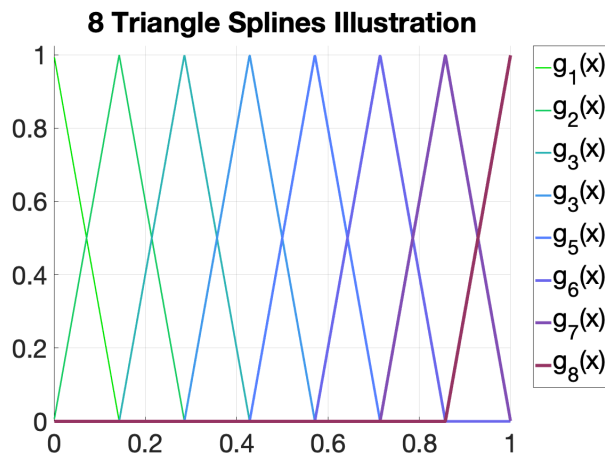


Figure 1: The figure illustrates our set of eight triangle splines in each individual coordinate dimension. Each is zero for all coordinates outside the base of its triangle. Our tensor spline is comprised of all products of the eight vertical and eight horizontal coordinate splines.

Table 2 presents average 95% confidence interval lengths for our three HAC bandwidths and HR for regressions that include our 8 by 8 set of spline basis terms. The format of rows displaying results for

$\rho$	Corr	No Splines				Triangle Splines			
		HAC $2\sigma$				HAC $2\sigma$			
		.05	.10	.15	HR	.05	.10	.15	HR
0.0	0.00	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
0.1	0.05	0.07	0.07	0.07	0.08	0.05	0.05	0.05	0.06
0.2	0.10	0.13	0.12	0.11	0.14	0.05	0.05	0.05	0.06
0.3	0.15	0.16	0.13	0.12	0.18	0.04	0.04	0.05	0.04
0.4	0.20	0.24	0.19	0.16	0.27	0.06	0.05	0.06	0.06
0.5	0.25	0.26	0.21	0.16	0.32	0.06	0.06	0.06	0.06
0.6	0.30	0.30	0.22	0.17	0.37	0.06	0.06	0.06	0.07
0.7	0.35	0.37	0.28	0.22	0.48	0.07	0.07	0.07	0.08
0.8	0.39	0.39	0.28	0.23	0.52	0.09	0.07	0.07	0.09
0.9	0.44	0.43	0.31	0.24	0.57	0.09	0.08	0.07	0.11
1.0	0.49	0.42	0.30	0.22	0.59	0.13	0.11	0.10	0.18

Table 1: Rejection frequencies testing the true null hypothesis of zero slope with nominal 5% t-tests for different levels of spatial correlation ( $\rho$ ). HAC estimates use Gaussian kernels with  $2\sigma = .05, .10, .15$ . Right panel uses tensor product of 8 triangle splines illustrated in Figure 1. Column labeled ‘Corr’ displays correlation of points at distance of .1. 1000 simulations.

$\rho$	Corr	HAC Bwidth $2\sigma$			
		.05	.10	.15	HR
0.0	0.00	0.19	0.19	0.19	0.19
0.1	0.05	0.19	0.19	0.19	0.19
0.2	0.10	0.19	0.19	0.19	0.19
0.3	0.15	0.19	0.19	0.19	0.19
0.4	0.20	0.19	0.19	0.19	0.19
0.5	0.25	0.19	0.19	0.19	0.19
0.6	0.30	0.19	0.19	0.19	0.19
0.7	0.35	0.19	0.19	0.20	0.19
0.8	0.39	0.20	0.20	0.20	0.19
0.9	0.44	0.20	0.21	0.22	0.19
1.0	0.49	0.22	0.23	0.24	0.19

Table 2: Nominal 95% Confidence Interval length for differing DGPs and different HAC estimators. HAC estimates use Gaussian kernels with  $2\sigma = .05, .10, .15$ . Spatial basis is an 8x8 tensor of triangular B-splines. Column labeled ‘Corr’ displays correlation of points at distance of .1. 1000 simulations.

differing values of  $\rho$  is analogous to Table 1. Entries are averages across 1000 simulations of nominal 95% confidence intervals.

The HR confidence intervals have average length about .19. HAC confidence interval lengths for smaller values of  $\rho$  are also about .19 and then slowly increase as  $\rho$  increases until about .20 at  $\rho = .8$ . HAC coverage probabilities remain fairly accurate for  $\rho$  between 0 and .8 without a large increase in their average length. For example, with a bandwidth of  $2\sigma = .1$  there is at most a 2% size distortion, nominal 95% intervals cover at 93%. With our approach these intervals are both close to nominal coverage and remain short enough to be scientifically useful. Even with the two most extreme correlation levels  $\rho = .9, 1$  in the Table, the intervals do not explode in length with averages of .20 to .24 across bandwidths. This paired with size distortions of at most 8% and only 5% with the largest bandwidth imply these intervals perform well even with very high levels of spatial dependence.

Figure 2 presents five sub-graphs illustrating the performance of our spatial basis pre-whitening approach. These figures display results from 1000 simulations of the mixture process described above for  $\rho = .8$ . In each simulation, using the real locations, 500 observations of  $X_i$  and  $Y_i$  are generated. We then estimate an OLS regression of  $Y_i$  on  $X_i$  and  $G_i$ , for a variety of specifications of  $G_i$ . The various  $G_i$  specifications are all constructed based upon an 8 by 8 tensor product of triangle B-splines evaluated at the real locations. First the 64 principal components (PCs) of this tensor product are computed. Then options for  $G_i$  are taken as the first PC, the first two PCs, first three PCs, and so on until all 64 PCs are used. The horizontal axis in each subgraph indicates how many PCs were used for  $G_i$ , thus reading the graphs from left to right illustrates how results change as the number of PCs is increased.

These sub-graphs simply present averages across simulations of characteristics of a set of fixed models. The next Section will investigate the performance of model selection algorithms that may choose data-dependent  $G_i$  specifications across simulations and thereby improve inference procedures.

The sub-graph labeled ‘HAC  $2\sigma = .10$  Reject’ presents rejection frequencies for a set of nominal 5% t-tests of the true null hypothesis of zero slope using a Gaussian kernel HAC estimator with two standard deviation ‘bandwidth’ equal to .10. As the number of PCs increase, the rejection frequencies generally decline and approach 7% when all 64 PCs are the constituents of  $G_i$ . Comparing these rejection frequencies to the 28% rejections reported in Table 1 for the corresponding HAC estimator without a spatial basis reveals a very substantial improvement in size as the number of PCs is increased.

The sub-graph labeled ‘Avg. CI’ presents the average 95% Confidence Interval length across simulations. As the number of terms in  $G_i$  grows, initially these average confidence intervals shrink in length even as their coverage properties improve. Eventually, as the number of PCs climbs above 50 the average CI length begins to rise slowly. When all PCs are used it is approximately 3% larger than it’s minimum length. This is in line with the anticipated effects of increasing the number of terms in the spatial basis  $G_i$ . Adding terms will reduce spatial correlation in residuals which will tend to lower the variance of the  $\hat{\beta}$  estimator but it will also remove some of the identifying variation in  $X_i$  which acts to increase the variance of  $\hat{\beta}$ . It appears that the first effect dominates up to about 40-50 PCs and after that the latter dominates.

## Properties of Alternate G Specifications Using Principal Components

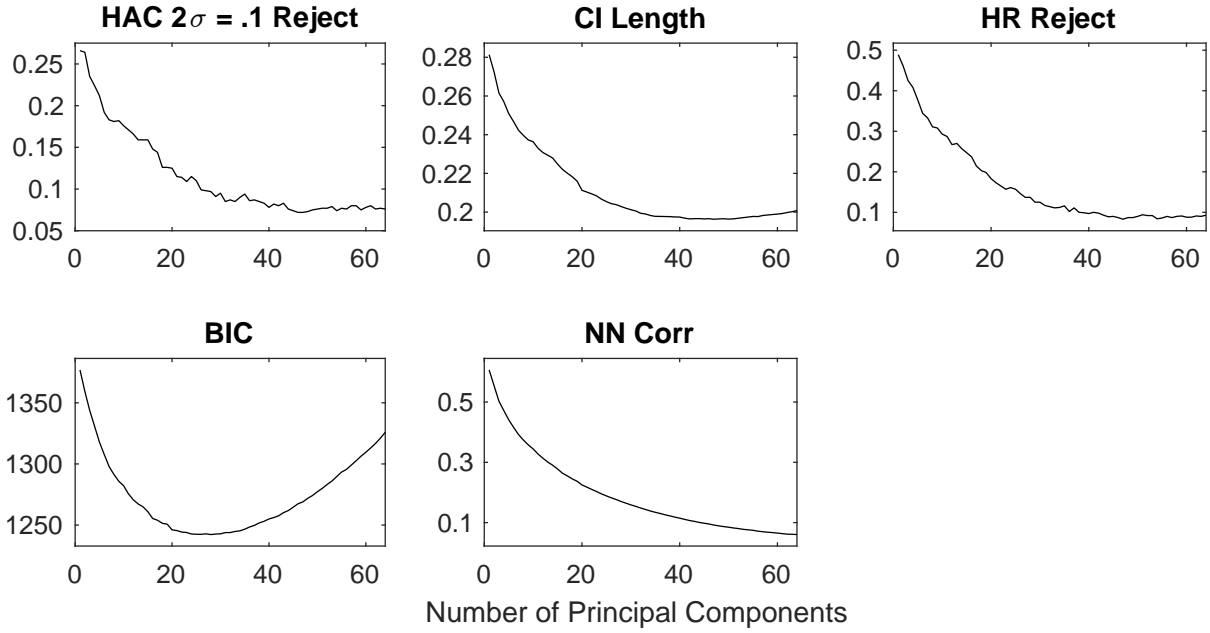


Figure 2: The horizontal axis indexes the number of principal components from an 8 by 8 tensor product of triangle B-splines used in the spatial basis,  $G_i$ . The DGP uses  $\rho = .8$ .

The sub-graph labeled ‘HR Reject’ displays rejection frequencies for heteroskedasticity robust standard errors, with no spatial dependence correction. For small numbers of PCs there are unsurprisingly very large size distortions. However, as the number of PCs approaches 64 these rejection frequencies approach about 9%, the spatial basis drastically reduces the spatial dependence in scores.

The second row of sub-graphs illustrate potential model selection criteria, Bayesian Information Criteria (BIC) and nearest neighbour correlations in residuals, labeled ‘BIC’ and ‘NN Corr’ respectively. In interpreting the BIC sub-graph recall that averages across simulations for a given number of PCs are displayed, not the results of a search for minimum BIC within each simulation. This graph still illustrates a tendency for BIC to be lower with intermediate numbers of PCs and then rise as the number of PCs approach 64. Nearest neighbor correlations in contrast have a tendency to decline as the number of PCs increases. We examine two candidate data-driven  $G_i$  choice methods in the following Section.

### 5. How to Pick $G_i$ ? Simulation evidence for two methods

In this Section, we examine potential methods for choosing  $G_i$  when the data have two-dimensional coordinates and the functions that underlie the construction of  $G_i$  are triangle B-splines evaluated at the coordinates. We investigate using either (1) an absolute nearest neighbor (NN) correlation in residuals to choose  $G_i$  and pair it an ad hoc HAC bandwidth or (2) use a method due to [Cao et al., 2023] that uses a simulation exercise to choose a combination of kernel bandwidth,  $G_i$ , and critical values. We refer to this as the CHKV method.

With both methods, candidates for  $G_i$  are sets of principal components of matrices of tensor products of B-splines. First, we construct a tensor products of B-splines in each dimension and calculate its principal components (PCs). The set of potential candidates for  $G_i$  is the collection consisting of  $G_i$  corresponding to the first PC,  $G_i$  corresponding to the first two PCs, and so on until  $G_i$  corresponding to the full set of PCs. In our main simulations, we examine tensor products of 10 triangle B-splines in each dimension to form the set of candidate  $G_i$ . Results for 8 by 8 tensors are in the Appendix.

Our simulations described below use the same locations, set of DGPs, and simulation sample sizes as in the previous Section.

### Nearest Neighbor Residual Correlation

For each simulated dataset, we estimate our model via an OLS regression of simulated  $Y_i$  on simulated  $X_i$  and some  $G_i$ . For each candidate  $G_i$  we calculate the absolute value of the NN residual correlation. The  $G_i$  that minimizes this absolute NN residual correlation is the chosen model for that simulated dataset. We then compute a set of HAC estimates as in the previous section via Gaussian kernels with  $2\sigma = .05, .10, .15$ .

Tables 3 and 4 present simulation results for rejection frequencies and 95% confidence interval length for our NN selection method for G. Rejection frequencies in Table 3 are very similar to those in Table 1. There are drastic size improvements versus not using splines, size distortions are small, 1% to 2% up to  $\rho = .8$  with little variation across the bandwidths here. Confidence interval lengths in Table 4 are also similar to those in Table 1, length does increase with  $\rho$  but not drastically. For lower levels of  $\rho$  there appears a slight advantage of the NN method in generating shorter average length confidence intervals but they are still close, often .18 versus .19. Even for the most severe levels of dependence, size distortions are modest with the larger bandwidth HAC estimator, at most 4%.

### CHKV Simulation-based tuning parameters

The CHKV approach uses an approximate model DGP  $F(\tau)$  to conduct, within each simulation, a Monte Carlo exercise to select a  $G_i$ , a bandwidth for a Gaussian kernel HAC standard error estimator, and critical values. Refer to a given combination of these as  $(G_i, bw, cv)$ . The algorithm will search over gridded-up set of combinations of G and bw.<sup>6</sup> The procedure for each simulation is: (1) fit the approximate model to the simulated dataset to get an approximate DGP:  $F(\hat{\tau})$ . (2) Using the real locations, generate a large sample of Monte Carlo (MC) draws from  $F(\hat{\tau})$ . (3) For each candidate  $(G_i, bw)$  use the MC draws to determine a 5% critical value for a t-test of the true null of zero slope  $cv_{MC}$  as the 95th percentile of the MC absolute t-statistics. (4) Take as the critical value,  $cv$ , for the triple  $(G, bw, cv)$  the  $\max(cv_{MC}, 1.96)$ . (5) Use the MC draws to estimate average power versus a set of false null hypotheses regarding the regression slope.<sup>7</sup> This will create an average power number for each  $(G_i, bw, cv)$  option. (6) Choose the  $(G_i, bw, cv)$  that has the highest average power and use it for the simulation dataset.

<sup>6</sup>This grid contained all combinations of Gaussian Kernel bandwidth ( $2\sigma$ ) taking values of [0.0, 0.025, 0.05, 0.075, 0.10, 0.125, 0.15] and number of PCs of the 10x10 B-splines from the set [0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 99].

<sup>7</sup>We calculated power for 2 alternatives, slope =  $\pm 3/N^{1/2}$

$\rho$	Corr	No Splines				Triangle Splines					
		HAC				HAC					
		.05	.10	.15	HR	.05	.10	.15	HR	NN	PCs
0.0	0.00	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	-0.02	11.55
0.1	0.05	0.07	0.07	0.07	0.08	0.05	0.05	0.06	0.05	0.00	21.52
0.2	0.10	0.13	0.12	0.11	0.14	0.07	0.07	0.07	0.07	0.00	32.58
0.3	0.15	0.16	0.13	0.12	0.18	0.05	0.05	0.06	0.05	0.00	40.02
0.4	0.20	0.24	0.19	0.16	0.27	0.06	0.06	0.06	0.06	0.00	48.72
0.5	0.25	0.26	0.21	0.16	0.32	0.05	0.05	0.06	0.06	0.00	60.26
0.6	0.30	0.30	0.22	0.17	0.37	0.06	0.06	0.06	0.06	0.00	70.48
0.7	0.35	0.37	0.28	0.22	0.48	0.07	0.07	0.07	0.08	0.01	82.56
0.8	0.39	0.39	0.28	0.23	0.52	0.06	0.06	0.06	0.06	0.01	93.45
0.9	0.44	0.43	0.31	0.24	0.57	0.09	0.07	0.06	0.10	0.03	97.29
1.0	0.49	0.42	0.30	0.22	0.59	0.11	0.10	0.09	0.15	0.05	97.69

Table 3: Rejection frequencies testing the true null hypothesis of zero slope at the nominal 5% significance level for different values of spatial dependence indexed by  $(\rho)$ . Gaussian kernel HAC variance estimators use  $2\sigma = .05, .10, .15$ . For each simulation, the spatial basis consists of some number of principal components of a **10x10** tensor of **triangle** B-splines. The number of PCs is chosen to minimize the absolute value of residuals' nearest neighbor correlation. The column labeled NN reports the average nearest neighbor residual correlation across simulations. PCs is the average number of principal components used across simulations. Column *Corr* shows the DGP correlation at distance  $h = 0.1$ ,  $Corr = \rho * \exp(-\frac{1}{\sqrt{2}})$ . 1000 simulations.

Table 5 presents simulation results for rejection frequencies and 95% confidence interval length for the CHKV. This Table shows a very different pattern in rejection frequencies as  $\rho$  varies compared with Tables 1 and 3. There are minimal size distortions for high and low  $\rho$  but distortions of up to 8% for medium values. Because the CHKV method uses simulations from an approximate model  $F(\tau)$  to choose bandwidth,  $G_i$ , and critical values, it is vulnerable to the choice of  $F(\tau)$ . Our simulation DGPs have a discontinuity in the covariance function at zero when  $\rho$  is not 1 or 0. The  $F(\tau)$  we use for the CHKV procedure does not allow such a discontinuity at zero, its covariances decay geometrically and are continuous at zero. This causes  $F(\tau)$  to be a better approximation of the true simulation DGP for high and low  $\rho$  than for medium  $\rho$ . For  $\rho = 0, 1$  it nests the true DGP. The CHKV results are still a vast improvement versus not using our method at all but their reliance on high quality is a drawback.

$\rho$	Corr	HAC			HR
		.05	.10	.15	
0.0	0.00	0.18	0.18	0.17	0.17
0.1	0.05	0.18	0.18	0.18	0.17
0.2	0.10	0.18	0.18	0.18	0.18
0.3	0.15	0.18	0.18	0.18	0.18
0.4	0.20	0.19	0.19	0.19	0.17
0.5	0.25	0.19	0.19	0.19	0.17
0.6	0.30	0.19	0.19	0.20	0.17
0.7	0.35	0.20	0.20	0.20	0.17
0.8	0.39	0.20	0.21	0.21	0.17
0.9	0.44	0.21	0.21	0.22	0.17
1.0	0.49	0.22	0.23	0.24	0.17

Table 4: 95% Confidence Interval length of different HAC variance estimators. Gaussian kernel HAC variance estimators use  $2\sigma = .05, .10, .15$ . The number of principal components of a  $10 \times 10$  tensor of triangular B-splines that minimize the residuals' absolute nearest neighbor correlation. 1000 simulations. Column *Corr* shows the theoretical correlation at distance  $h = 0.1$ ,  $corr = \rho * \exp(-\frac{1}{\sqrt{2}})$ .

Table 5 shows the rejection frequencies and average CI lengths for 1000 simulations using the CHKV method to choose (G, bw, cv). The rows of the Table refer to differing values of  $\rho$  in our simulation DGP, just as in previous tables. Likewise the second column provides the true DGP correlation at distance of .1. Ref. F. reports rejection frequencies of the true null hypothesis of zero slope. The last two columns report average 95% length and the average number of PCs used across the simulations. A 10 by 10 tensor of triangle B-splines is used.

$\rho$	Corr.	Rej. F.	CI Length	Avg PCs
0.0	0.00	0.04	0.18	7.54
0.1	0.05	0.05	0.18	3.35
0.2	0.10	0.07	0.18	1.98
0.3	0.15	0.12	0.18	3.86
0.4	0.20	0.12	0.19	10.76
0.5	0.25	0.12	0.20	25.29
0.6	0.30	0.09	0.20	43.83
0.7	0.35	0.06	0.21	59.39
0.8	0.39	0.04	0.22	70.15
0.9	0.44	0.04	0.23	76.19
1.0	0.49	0.05	0.26	84.58

Table 5: Rejection frequencies and confidence Interval lengths using the bandwidth/ $G_i$  selection method. Approximate covariance matrix is  $\exp(\tau_0) \exp(-\tau_1 \cdot D)$ , where  $D$  is the matrix of the euclidean distances between the locations. 10 by 10 tensor product of spline.

## 6. Empirical Example Application

To illustrate our method we apply it to a regression from the study, “Pre-Colonial Ethnic Institutions and Contemporary African Development” [Michalopoulos and Papaioannou, 2013], first column of Table 2.1. The specification is a bi-variate regression of log night-time illumination of an ethnic group’s homeland region on a treatment variable that is the degree of centralization of the governance of the group’s historical tribe ranging from 0 for stateless to 3 for strong centralized states. The regions/groups are derived from on Murdoch’s “Ethnographic Atlas”. Original study standard errors used HAC [Conley, 1999] and clustered by country and ethnic group. The study’s 683 observations are plotted in Figure 5 below.

Figure 3 illustrates our NN objective function for different size tensors, from 9 by 9 to 12 by 12. The 12 by 12 tensor yeilds a NN correlation of approximately zero with the fewest number of PCs so that is our baseline choice for  $G_i$ . Our NN approach uses the information in Figure 3 to choose a  $G_i$ . We anticipate that many researchers will want to consider tradeoffs in NN residual correlation with the number of variables in  $G_i$ . Since we are using HAC standard errors, inference using  $G_i$  with NN correlations that are small may work just as well as with a larger  $G_i$  with NN correlations of zero. Based on the Figure, we will examine  $G_i$  choices of 65 and 50 PCs from a 12 by 12 tensor with NN correlations of approximately zero and 5% respectively.

Tables 6 and 7 present point estimates and standard errors of the treatment slope of the original MP regression and our estimates for different numbers of PCs and alternate HAC estimators. Here, we differ from our simulation sections and use a uniform kernel for our HAC estimators for ease in interpretation. In Table 6, 65 PCs yeilds a NN residual correlation of approximately zero while 50 results in about a 5% NN correlation. Standard errors in both Tables are very similar to each other and across HAC bandwidths. Our standard error estimates are much smaller than those in MP which did attempt to

	Original	HAC(150)	HAC(250)	HAC(350)
Point Estimate	.41	.23	.23	.23
Std Error	.12	.07	.06	.07

Table 6: Point Estimates and Standard Errors, Original and HAC with bandwidths of 150, 250, and 350km. G 65 PCs from 12 by 12 tensor product.

	Original	HAC(150)	HAC(250)	HAC(350)
Point Estimate	.41	.22	.22	.22
Std Error	.12	.07	.07	.07

Table 7: Point Estimates and Standard Errors, Original and HAC with bandwidths of 150,250, and 350km. G 50 PCs from 12 by 12 tensor product.

allow for dependence with two-way cluster and with HAC standard errors (which turned out to be very similar). It is also important to note the shift in our point estimates versus MP. Overall, our confidence intervals are substantially lower and shorter.

Absolute Correlation between Nearest Residuals

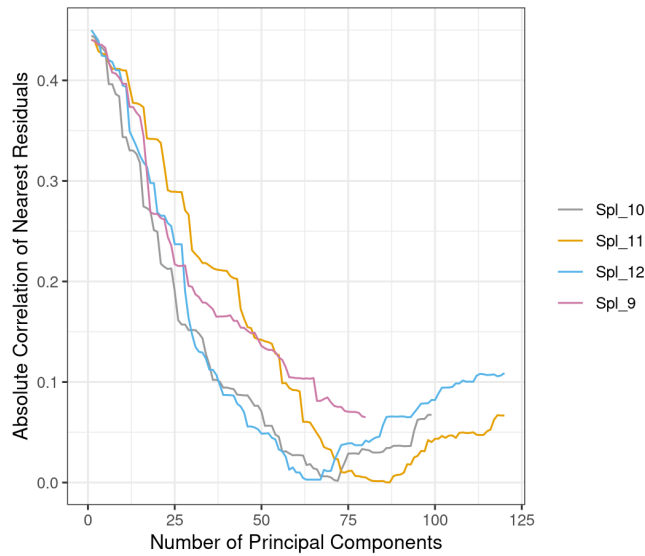


Figure 3: Absolute NN correlations for choices of tensor from 9 by 9 to 12 by 12 and alternate numbers of PCs.

Figure 5 displays the estimates of our “surface” using 60 PCs from a 12 by 12 tensor product of triangle B-splines. South Africa is in the foreground of the plot.

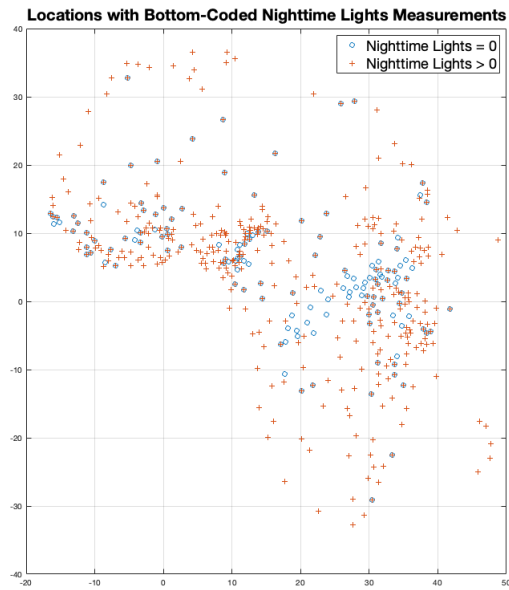


Figure 4: The Figure illustrates locations from the data used in [Michalopoulos and Papaioannou, 2013] in which nighttime lights measurements are sufficiently low to be bottom coded.

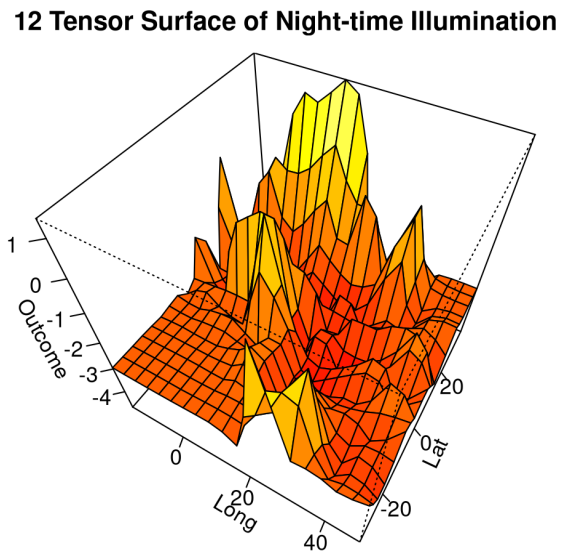


Figure 5: The Figure illustrates estimates  $G\hat{\gamma}$  versus coordinates as an estimated "surface" using 60 PCs from a 12 by 12 tensor product of triangle B-splines.

## References

- [Andrews, 1991] Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858.
- [Assoud, 1977] Assoud, P. (1977). *Espaces Métriques, Plongements, Facteurs*. Doctoral Dissertation, Université de Paris XI, 91405 Orsay France.
- [Bartlett, 1950] Bartlett, M. S. (1950). Periodogram analysis and continuous spectra. *Biometrika*, 37:1–16.
- [Bester et al., 2011a] Bester, C. A., Conley, T. G., and Hansen, C. B. (2011a). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, 165(2):137 – 151.
- [Bester et al., 2011b] Bester, C. A., Conley, T. G., and Hansen, C. B. (2011b). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, 165(2):137–151.
- [Cao et al., 2023] Cao, j., Hansen, C., Kozbur, D., and Villacorta, L. (Forthcoming, 2023). Inference for dependent data with learned clusters. *Review of Economics and Statistics*.
- [Conley, 1996] Conley, T. G. (1996). *Econometric Modelling of Cross-Sectional Dependence*. Ph.D. Dissertation, University of Chicago.
- [Conley, 1999] Conley, T. G. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics*, 92:1–45.
- [Conley et al., 2023] Conley, T. G., Goncalves, S., Kim, M. S., and Perron, B. (2023). Bootstrap inference under cross-sectional dependence. *Quantitative Economics*, 14(2):511–569.
- [DellaVigna et al., 2025] DellaVigna, S., Imbens, G., Kim, W., and Ritzwoller, D. (2025). Using multiple outcomes to adjust standard errors for spatial correlation. *Working Paper*.
- [Ibragimov and Müller, 2010] Ibragimov, R. and Müller, U. K. (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics*, 28(4):453–468.
- [Jenish and Prucha, 2009] Jenish, N. and Prucha, I. R. (2009). Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of Econometrics*, 150(1):86–98.
- [Lazarus et al., 2018] Lazarus, E., Lewis, D. J., Stock, J. H., and Watson, M. W. (2018). HAR Inference: Recommendations for Practice. *Journal of Business & Economic Statistics*, 36(4):541–559.
- [Michalopoulos and Papaioannou, 2013] Michalopoulos, S. and Papaioannou, E. (2013). Precolonial ethnic institutions. *Econometrica*, 81:113–152.
- [Müller and Watson, 2024] Müller, U. and Watson, M. (2024). Spatial unit roots and spurious regression. *Working Paper*.
- [Müller and Watson, 2022] Müller, U. K. and Watson, M. W. (2022). Spatial correlation robust inference. *Econometrica*, 90(6):2901–2935.
- [Stein, 1972] Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Sixth Berkely Symposium*, pages 583–602.

[Sun and Kim, 2015] Sun, Y. and Kim, M. S. (2015). Asymptotic  $F$ -test in a GMM framework with cross-sectional dependence. *Review of Economics and Statistics*, 97(1):210–223.

APPENDIX

Additional results with 8 by 8 tensors for NN and CHKV approaches.

$\rho$	Corr	No Splines				Triangle Splines					
		HAC				HAC				NN	PCs
		.05	.10	.15	HR	.05	.10	.15	HR		
0.0	0.00	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	-0.02	11.64
0.1	0.05	0.07	0.07	0.07	0.08	0.06	0.05	0.06	0.06	0.00	21.10
0.2	0.10	0.13	0.12	0.11	0.14	0.07	0.07	0.07	0.07	0.00	31.39
0.3	0.15	0.16	0.13	0.12	0.18	0.05	0.05	0.05	0.05	0.00	39.07
0.4	0.20	0.24	0.19	0.16	0.27	0.06	0.05	0.06	0.06	0.01	46.61
0.5	0.25	0.26	0.21	0.16	0.32	0.06	0.06	0.06	0.06	0.01	54.73
0.6	0.30	0.30	0.22	0.17	0.37	0.07	0.06	0.06	0.07	0.02	59.83
0.7	0.35	0.37	0.28	0.22	0.48	0.07	0.07	0.07	0.08	0.04	62.51
0.8	0.39	0.39	0.28	0.23	0.52	0.09	0.07	0.07	0.09	0.06	63.14
0.9	0.44	0.43	0.31	0.24	0.57	0.09	0.08	0.07	0.11	0.08	63.16
1.0	0.49	0.42	0.30	0.22	0.59	0.13	0.11	0.10	0.18	0.10	63.32

Table 8: Rejection frequencies testing the true null hypothesis of zero slope, at the 5% significance level for different values of spatial correlation ( $\rho$ ). Gaussian kernel HAC variance estimators use  $2\sigma = .05, .10, .15$ . For each simulation the number of principal components of the **8x8 triangle** B-splines is chose to minimize the residual's absolute nearest neighbor correlation. The column labeled NN presents the average absolute nearest neighbor correlation and the colum labeled PCs presents the average number of chosen PC terms across simulations. Column *Corr* shows the DGP correlation at distance .1 for each value of  $\rho$ . 1000 simulations.

$\rho$	Corr	HAC			HR
		.05	.10	.15	
0.0	0.00	0.18	0.18	0.17	0.17
0.1	0.05	0.18	0.18	0.18	0.17
0.2	0.10	0.18	0.18	0.18	0.18
0.3	0.15	0.18	0.18	0.18	0.18
0.4	0.20	0.18	0.19	0.19	0.17
0.5	0.25	0.19	0.19	0.19	0.17
0.6	0.30	0.19	0.19	0.19	0.17
0.7	0.35	0.19	0.19	0.20	0.17
0.8	0.39	0.19	0.20	0.20	0.17
0.9	0.44	0.20	0.21	0.22	0.17
1.0	0.49	0.22	0.23	0.24	0.17

Table 9: Confidence Interval length of different HAC variance estimators for the standard error. Gaussian kernel HAC variance estimators use  $2\sigma = .05, .10, .15$ . The number of principal components of the **8x8 triangular** B-splines minimizes the residuals' absolute nearest neighbor correlation. 1000 simulations. **500** points. Column *Corr* shows the theoretical correlation at distance  $h = 0.1$ ,  $corr = \rho * \exp(-\frac{1}{\sqrt{2}})$ .