Improved Spatial Dependence-Robust Inference via Pre-whitening

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

September 16, 2024

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.[1] 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.[2] 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

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

[1] At least since [Conley, 1996] and [Conley, 1999]

[2] See for time series [Bartlett, 1950], [Andrews, 1991].

[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 require a worst-case scenario assumption regarding covariance functions and an assumption regarding the extent of spatial covariances.

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.

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 and they are mixing and obey a standard set of regularity conditions so a law of large numbers and central limit theorem apply. Close ob-

servations 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]

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. Future drafts of this paper will include a Section 5 which does two things. First it will examine Monte Carlo performance of simple algorithms to implement the choice of spatial basis terms. We consider using either a BIC penalty and a nearest neighbor correlation criteria for model

selection. In addition, we will include Section 6 which will contain a Monte Carlo evaluation of using a parametric approximating model to construct critical values following [Cao et al., 2023].

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 $d : S \times S \to [0, \infty)$. 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 assume that the data are weakly dependent and that observations i and j approach independence as d_{ij} grows large.

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

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

The random variables ε_i are unobserved and β is identified through the usual conditions that $\mathrm{E}[\varepsilon_i X_i] = 0$ and $\mathrm{E}[X_i X_i']$ is full rank. With weakly dependent data, Ordinary Least Squares (OLS) estimates of β_0 are consistent.

Given that consistent estimates can be constructed, an accompanying problem is constructing a $1 - \alpha$ level confidence set $\widehat{C}$ for β_0, that satisfies

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

where ν_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., $\Pr(\widehat{C}$ contains $\beta_0)$ may in practice be far from $1 - \alpha$.) Standard methods for constructing $\widehat{C}$ in the context of the linear model with sufficiently strongly mixing properties for observations across i, is to estimate $\widehat{\beta}$ 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 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 generated from what we term a spatial pre-whitening basis. A spatial pre-whitening basis is a set $\mathcal{G}$ of functions $g \in \mathcal{G}$ of the spatial indexing set, each of the form $g : S \to [0, 1]$. Then $G_i \in \mathbb{R}^{\mathcal{G}}$ is defined with components $[G_i]_g = g(i)$. The main examples of spatial pre-whitening bases that we discuss below are spatially localized B-splines.

Our procedure is defined as follows. To construct a confidence interval for a component of β_0, first estimate $[\widehat{\beta}, \widehat{\gamma}]$ with OLS regression Y_i on $[X_i, G_i]$. Subsequently construct the usual confidence interval using spatial HAC estimation with bandwidth $h > 0$ and kernel function k for the above regression. Let $\widehat{V}_{(k,h)}$ be the corresponding estimate. For a component $[\beta_0]_j$ of interest of β_0, let q_a be the ath quantile of the standard Gaussian random variable (i.e., of $\mathrm{N}(0, 1)$), and set the total margin of error estimate $\widehat{\text{m.e.}} = q_{1-\alpha/2}[\widehat{V}_{(k,h)}]_{jj}^{-1/2}$.

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

More generally, confidence sets for functionals $a(\beta_0)$ may be constructed using the delta method the usual way. Confidence ellipsoids covering β_0 are also constructed using the usual asymptotic Gaussian approximation. Note that $\widehat{C} = \widehat{C}_j \times \mathbb{R}^{p\setminus\{j\}}$ covers all of β_0 with the same probability as $\widehat{C}_j$ 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 $\widehat{C}$ 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 $\mathrm{E}[\varepsilon_i | X_j] = 0$ 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 $v \in [0,1]$ depending on two arguments, $|\mathrm{E}[(v(Z_A, Z_B)] - \mathrm{E}[v(Z_A, Z'_B)]| \leq 2\exp(-d_{AB}/L_{\text{mix}})$.
4. (*Conditioning.*) For $R \subseteq S$, $\lambda_{\min}(|R|^{-1} \sum_{i \in R} \mathrm{E}[X_i X_i']) \geq 1/L_{\text{cond}}$ and $\lambda_{\min}(|R|^{-1}\mathrm{E}[\left(\sum_{i \in R} \varepsilon_i X_i\right)\left(\sum_{i \in R} \varepsilon_i X_i\right)']) \geq 1/L_{\text{cond}}$.
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 $g : S \to [0, 1]$. 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_{growth} in that $|1 - K(d)| \leq L_{\text{kern}} x^{L_{\text{growth}}}$. 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 $1 \leq (\ell_{basis})_n |\text{supp}(g)| \leq (\ell_{basis})_n L_{\text{basis}} |\{i \in S : |\tilde{g}(i)| > 1/L_{\text{basis}}\}|$. 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_{growth} is a characteristic that Euclidean spaces do also have.[3] If S is part of a sequence of cubes in an integer lattice, then L_{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

[3] 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_{growth}.

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 $x \in [0, 1]$.

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 ν which depends only on F which satisfies $\lim_{n\to\infty} \nu_n = 0$ 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 an asymptotic coverage of $1 - \alpha$, up to a remainder term ν_n which vanishes under $n \to \infty$. 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 asymptically 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 like [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 β_0, $p = 1$. The case $p > 1$ is analogous, noting that p is implicitly restricted by $(L_{\text{mom}}, L_{\text{cond}})$.

Let $\ell = \max((\ell_{\text{kern}})_n, (\ell_{\text{basis}})_n)$. 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 $|B_x(i)| \leq L_{\text{growth}} |B_{x/2}(i)| \leq L_{\text{growth}}^2 |B_{x/4}(i)| \leq ... \leq L_{\text{growth}}^{\lceil \log x \rceil + 1} |\{i\}|$, 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

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

Proof of Lemma 2. By Lemma 1, $|\Delta| \leq |T|L^{\log x+2}$. Then by partitioning the following sum, $\text{E}[(|T|^{-1}(\sum_{i\in T} W_i - \text{E}[W_i])^2] = |T|^{-2}\text{E}[\sum_{i\in\Delta}(W_{i_1} - \text{E}[W_{i_1}])(W_{i_2} - \text{E}[W_{i_2}]) + \sum_{i\in T^2\setminus\Delta}(W_{i_1} - \text{E}[W_{i_1}])(W_{i_2} - \text{E}[W_{i_2}]) \leq |T|^{-2}(|\Delta|(2L)^2 + |T|^2(2L)^2 2\exp(-(x-\nu)/L))$. 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 $\widehat{\xi}, \widehat{\eta}$ and $\widehat{\zeta}$, which are defined as least squares coefficients X_i, ε_i and Y_i on G_i. Denote $\tilde{X}_i = X_i - G_i\widehat{\xi}$, $\tilde{\varepsilon}_i = \varepsilon_i - G_i\widehat{\eta}$ and $\tilde{Y}_i = Y_i - G_i\widehat{\zeta}$.

Lemma 3. For every $g \in \mathcal{G}$, there is a set K_g with $\text{diam}(K_g) \leq \ell h$ such that $\widehat{\xi}_g$, $\widehat{\eta}$ and $\widehat{\zeta}_g$ 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. $\widehat{\xi}_g$ may be found by applying the Frisch Waugh Theorem. Then the least squares solution is $\widehat{\xi}_g = \left(\sum_{i\in\tilde{G}} \tilde{g}(i)^2\right)^{-1} \sum_{i\in\tilde{G}} X_i\tilde{g}(i)$. $\tilde{g}$ can also be defined using exclusively the $k \in \mathcal{G}$ with common points of support with g given by $K_g = \text{supp}(g) \cup \bigcup_{k:\text{supp}(k)\cap\text{supp}(g)\neq\varnothing} \text{supp}(k)$ and $\widehat{\xi}_g$ 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 $\widehat{\eta}$ and $\widehat{\zeta}$.

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,

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

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:

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

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

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

Proof of Lemma 5. To show the first statement suppose $i \in A, i \notin C_1 \cup C_2$. There must be π 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 πi_2 and πi_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 $f_2(R, x) = L^8 \exp(-(x/3 - 2\ell h)/L) + 4! |R|^{-2} L^{2 \log x + 12}$

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

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 $\max_{i \in A^\circ} \mathrm{E}[z_{i_1} z_{i_2} z_{i_3} z_{i_4}] \leq L \exp(-(x/3 - \ell h)/L)$. Simplifying and aggregating gives the proof.

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

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, $\mathrm{E}[(G_i' \widehat{\xi})^4] = \mathrm{E}[(\sum_{g : \mathrm{supp}(g) \ni i} g(i) \widehat{\xi}_g)^4] \leq |\{g : \mathrm{supp}(g) \ni i\}|^2 \sum_{g : \mathrm{supp}(g) \ni i} g(i)^4 \mathrm{E}[\widehat{\xi}_g^4] \leq L^7 \max_{g \in \mathcal{G}} f_2(\mathrm{supp}(\tilde{g}), x)$. Finally, $\tilde{X}_i = X_i - G_i \widehat{\xi}$ so $\mathrm{E}[\tilde{X}_i^4] \leq 4(L^4 + \max_{g \in \mathcal{G}} L^7 f_2(\mathrm{supp}(\tilde{g}), x))$.

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

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

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

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

Also needed is a central limit theorem for $\tilde{X}_i \varepsilon_i$, which is next. Let $S_{\boldsymbol{\ast}} \subseteq E$ be the image of a bi-Lipshitz Euclidean embedding ι 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

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

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 $\sigma^2 = \mathrm{var}(\Xi)$. Let $t \in \mathbb{R}$. Then

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

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

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 $x \ge 1$ and $c > 0$, $\Pr(|\Omega_0^K - \mathrm{E}[\Omega_0^K]| \ge c) \le c^{-2}f_4(x)$.

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 $\mathrm{E}[(\Omega_0^K - \mathrm{E}[\Omega_0^K])^2]$ expands to

$$\mathrm{E}\Big[\frac{1}{n^2}\sum_{i\in S^4} K_{i_1i_2}K_{i_3i_4}(z_{i_1}z_{i_2} - \mathrm{E}[z_{i_1}z_{i_2}])(z_{i_3}z_{i_4} - \mathrm{E}[z_{i_3}z_{i_4}])\Big]$$
$$= \frac{1}{n^2}\sum_{i\in A} K_{i_1i_2}K_{i_3i_4}\Big(\mathrm{E}[z_{i_1}z_{i_2}z_{i_3}z_{i_4}] - \mathrm{E}[z_{i_1}z_{i_2}]\mathrm{E}[z_{i_3}z_{i_4}]\Big).$$

Let $M_j = \max_{i\in C_j} K_{i_1i_2}K_{i_3i_4}|\mathrm{E}[z_{i_1}z_{i_2}z_{i_3}z_{i_4}] - \mathrm{E}[z_{\pi i_1}z_{\pi i_2}]\mathrm{E}[z_{\pi i_3}z_{\pi i_4}]|$, $j \le 3$. By $2\max_{i\in C_1} K_{i_1i_2}K_{i_3i_4}\mathrm{E}[|z_{i_1}z_{i_2}z_{i_3}z_{i_4}|] \le L$, $M_1 \le L$. Next decompose $M_2 \le$

$M_{2a}+M_{2b}$ with $M_{2a} = \max_{i\in C_2} K_{i_1i_2}K_{i_3i_4}|\mathrm{E}[z_{i_1}z_{i_2}z_{i_3}z_{i_4}]-\mathrm{E}[z_{\pi i_1}]\mathrm{E}[z_{\pi i_2}z_{\pi i_3}z_{\pi i_4}]|$, and $M_{2b} = \max_{i\in C_2} K_{i_1i_2}K_{i_3i_4}|\mathrm{E}[z_{\pi i_1}]\mathrm{E}[z_{\pi i_2}z_{\pi i_3}z_{\pi i_4}] - \mathrm{E}[z_{i_1}z_{i_2}]\mathrm{E}[z_{i_3}z_{i_4}]|$. 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, $\mathrm{E}[z_{\pi i_1}] = 0$ and either the bound $|\mathrm{E}[z_{i_1}z_{i_2}]| = |\mathrm{E}[z_{i_1}z_{i_2}] - \mathrm{E}[z_{i_1}]\mathrm{E}[z_{i_2}]| \leq L^4 \cdot L\exp(-(x-4\ell h)/L)$ 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 $|\mathrm{E}[z_{i_1}z_{i_2}z_{i_3}z_{i_4}] - \mathrm{E}[z_{i_1}z_{i_2}]\mathrm{E}[z_{i_3},z_{i_4}]| = |\mathrm{E}[z_{i_1}z_{i_2}z_{i_3}z_{i_4}] - \mathrm{E}[z_{\pi i_1}z_{\pi i_2}]\mathrm{E}[z_{\pi i_3},z_{\pi i_4}]| \leq L\exp(-(x-4\ell h)/L)$. If not, then either $d_{i_1i_2} \geq x \geq h$ or $d_{i_3i_4} \geq x \geq h$ and therefore $K_{i_1i_2} = 0$ or $K_{i_3i_4} = 0$. 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 $|C_2|, |C_3| \leq |A|$,

$$\mathrm{E}[(\bar{\Omega}_0^K - \mathrm{E}[\bar{\Omega}_0^K])^2] \leq \frac{1}{n^2}|C_1|M_1 + \frac{1}{n^2}|A|M_2 + \frac{1}{n^2}|A|M_3$$

$$\leq \frac{1}{n^2}\left(nL^{3\log 3x+6}L^8 + n^2L^{2\log h+4}3L^9\exp(-(x-4\ell h)/L)\right)$$

Using Markov's inequality and simplifying gives the lemma.

Next let $\delta_\beta = \beta_0 - \widehat{\beta}$. Denote $\mathrm{E}_S = n^{-1}\sum_{i_1\in S}$ and $\mathrm{E}_S^K = \sum_{i_2\in S} K_{i_1i_2}$. Then

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

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

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

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

Let

$$f_5(x) = 4L^6(1 + L\max_{g\in\mathcal{G}} f_2(\mathrm{supp}(\tilde{g}), x)L^{\log h+2})f_1(S, x, 4\ell h + 2h)$$

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

Lemma 11. For $j = 1, ..., 5$, $x \geq 1$ 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 δ_5, note $\widehat{\gamma}_g = (\sum_{i \in K_g} \tilde{g}(i)^2)^{-1} \sum_{i \in K_g} \tilde{g}(i) Y_i$. Let D_g be the denominator and $W_{i_1} = \sum_{i_2 \in B_h(i_1)} K_{i_1 i_2} \tilde{X}_{i_1} \varepsilon_{i_1} \tilde{X}_{i_2} G'_{i_2} \widehat{\gamma}$ so $\delta_5 = n^{-1} \sum_{i_1 \in S} W_{i_1}$ and

$$\mathrm{E}[W_{i_1}] = \mathrm{E}\left[\sum_{i_2 \in B_h(i_1)} K_{i_1 i_2} \bar{X}_{i_1} \varepsilon_{i_1} \tilde{X}_{i_2} \sum_{g \in \mathcal{G}} g(i_2) \widehat{\gamma}_g \right]$$

$$= \mathrm{E}\left[\sum_{i_2 \in B_h(i_1)} K_{i_1 i_2} \bar{X}_{i_1} \varepsilon_{i_1} \tilde{X}_{i_2} \sum_{g \in \mathcal{G}} g(i_2) D_g^{-1} \left(\sum_{i \in B_x(i_1)} \tilde{g}(i) Y_i + \sum_{i \in K_g \backslash B_x(i_1)} \tilde{g}(i) Y_i \right) \right].$$

For $i \in K_g \backslash B_x(i)$, 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 δ_3, note that $\widehat{\gamma} = \widehat{\xi} \delta_\beta + \widehat{\eta}$. Decompose

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

Formally replacing u for δ_β in the right hand term allows proceeding exactly as for δ_2 to calculate a bound. The left hand term is bounded exactly like δ_3.

The Lemma holds after simplifying.

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

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

Proof of Lemma 12. Use $T = S$ and Δ 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 $|\Delta| \times |1 - K_{i_1 i_2}| \leq nL_{\text{kern}} L_{\text{growth}}^2 L_{\text{growth}}^{\log x + xh^{-1} \log L_{\text{growth}}}$.

$$\mathrm{E}[\Omega_0^K] - \mathrm{E}[\Omega_0] = n^{-1} \sum_{i \in S^2} \mathrm{E}[(K_{i_1 i_2} - 1)\tilde{X}_{i_1} \varepsilon_{i_1} \tilde{X}_{i_2} \varepsilon_{i_2}]$$

$$= \sum_{R \in \{\Delta, S^2 \setminus \Delta\}} n^{-1} \sum_{i \in R} \mathrm{E}[(K_{i_1 i_2} - 1)\tilde{X}_{i_1} \varepsilon_{i_1} \tilde{X}_{i_2} \varepsilon_{i_2}].$$

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 here 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 have the same distribution and are independent of each other. Both variables have the same mixture distribution that combines idiosyncratic noise with a spatially correlated component resulting in a covariance matrix that is a linear combination of an identity matrix and a non-diagonal matrix Σ. So we generate variables X_i with:

$$X_i \sim N(0, 1),\ \mathrm{cov}(X_i, X_j) = (1 - \rho) + \rho \Sigma_{ij}$$

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

Table 1 presents results where Σ has parameter $\theta = \sqrt{2}/10$. To get better understand level of spatial correlation implied by this value of θ, 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 ρ 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 ρ 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 spline terms and different bandwidths using a Gaussian kernel, $N(0, \sigma^2 I)$. [4] The bandwidth is described by headings .05,.10, .15 which give the value of2σ for each kernel. The second panel uses the same HAC estimator but adds an 8×8 tensor product of triangular B-splines to the regression.[5]

Rows in Table 1 present differing values of ρ, starting from $\rho = 0$ when both X_i and Y_i are white noise. Subsequent rows present alternative mixture proportions. To illustrate the amount of correlation in both X_i and Y_i as ρ increases, the second column labeled 'corr' reports the correlation between pairs of obser-

[4] We highly recommend using a positive semi-definite kernel function $k(x)$. For larger numbers of spline terms, non-PSD kernels can yield negative variance estimates frequently enough to be an issue.

[5] 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.

vations 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 ρ 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 are apparent for ρ values of .2 and above. Size distortions become very severe as ρ 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 used in 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 ρ 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.

Table 2 presents average 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 differing values of ρ is analogous to Table 1. Entries are averages across simulations of nominal 95% confidence intervals.

The HR confidence intervals have average length about .19. HAC confidence interval lengths for smaller values of ρ are also about .19 and the slowly increase as ρ increases until about .20 at $\rho = .8$. HAC Coverage probabilities remain

		No Splines				Triangle Splines			
		HAC 2σ				HAC 2σ			
ρ	Corr	.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 (ρ). HAC esimates use Gaussian kernels with $2\sigma = .05, .10, .15$. Right panel uses tensor product of 8 triangle splines illustrated in Figure 1. Column labled 'Corr' displays correlation of points at distance of .1. 1000 simulations.

		HAC Bwidth 2σ			
ρ	Corr	.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: Confidence Interval length of different HAC variance estimators with an 8x8 tensor of triangular B-splines. 1000 simulations.

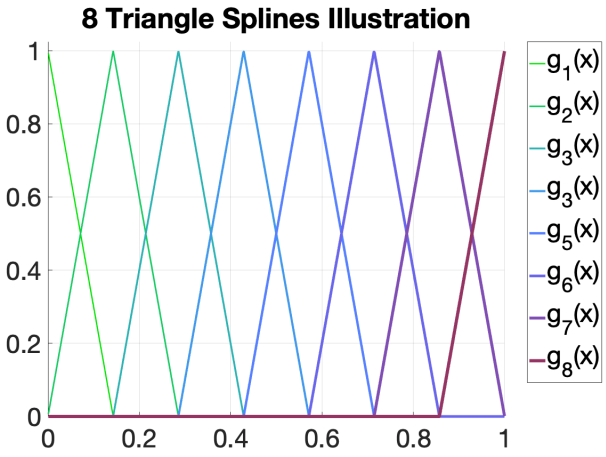

Figure 1: The figure illustrates our set of eight triangle splines in each individual coordinate dimension. Each is zero for all coordinates outside the base of its triangle. Our tensor spline is comprised of all products of the eight vertical and eight horizontal coordinate splines.

fairly accurate for ρ 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, very hign 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, 500 observations of X_i and Y_i are generated and an OLS regression of Y_i on X_i and a spatial basis G is conducted for a variety of specifications of $\mathcal{G}$. The various $\mathcal{G}$ are all constructed based upon and 8 by 8 tensor product of triangle B-splines. First the 64 principal components (PCs) of this tensor product are computed.

Then choices of G 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 subgraphs indicates how many PCs were used, 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 different $\mathcal{G}$ 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 included in $\mathcal{G}$. 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 $\mathcal{G}$ 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 $\mathcal{G}$. Adding terms will reduce spatial correlation in residuals which will tend to lower the variance of the $\widehat{\beta}$ estimator but it will also remove some of the identifying variation in X which acts to increase the variance of $\widehat{\beta}$. It appears that the first effect dominates up to about 40-50 PCs and after that the latter dominates.

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,

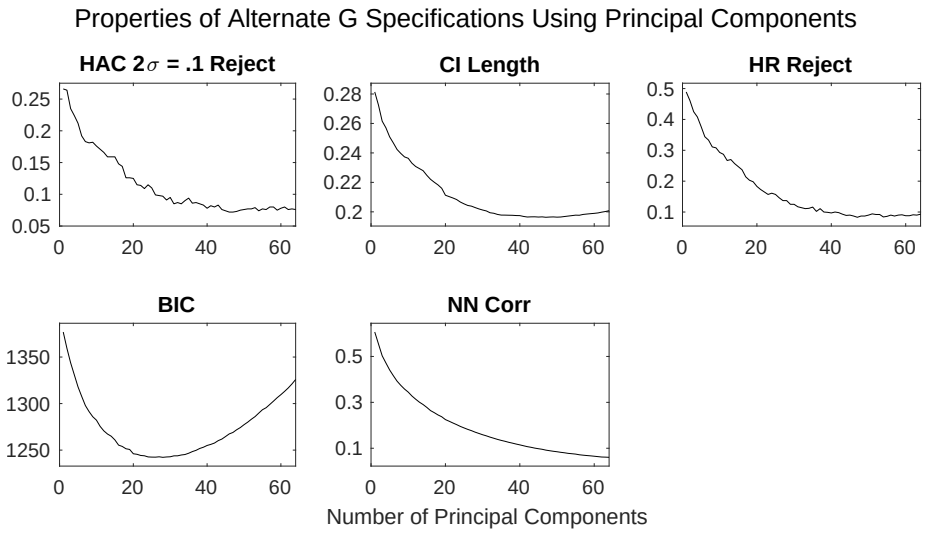

Figure 2: The horizontal axis indexes the number of principal components from an 8 by 8 tensor product of triangle B-splines used in the spatial basis, G. The DGP uses $\rho = .8$.

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. Therefore, we anticipate that choice of the components of $\mathcal{G}$ with differ across criteria based upon these two criteria. We examine two candidate $\mathcal{G}$ choice criteria in the following Section.

5. Pre-whitening Basis Selection

When this Section is completed, it will examine potential methods for choosing $\mathcal{G}$ when the data has a two-dimensional index and the functions used to construct $\mathcal{G}$ are triangle B-splines. We investigate using either BIC or an average nearest neighbor correlation in residuals as our model selection criteria. We examine two candidate sets of components for $\mathcal{G}$: sets of tensor products and sets of principal components of tensor products.

1. Alternate Sets of Full Tensors The first method is to construct a set of candidate $\mathcal{G}$s with each being a tensor product of triangle B-splines in each dimension. We then select from among these candidate full tensor products either via a BIC penalty or the one with nearest neighbor correlation nearest to zero. In our simulations, we consider tensor products of 4,5,6,7,8,9,10 triangle B-splines, thus resulting in a range of 16 to 100 terms in $\mathcal{G}$.

2. Principal Components of Tensors This method is to first construct a tensor products of B-splines in each dimension, $\mathcal{G}_0$, and calculate the principal components (PCs) of $\mathcal{G}_0$ to form $\mathcal{G}$. For each tensor we consider candidate $\mathcal{G}$ formed by the first PC, the first two PCs, and so on until the full set of PCs. In our simulations we examine tensor products of 4 to 10 triangle B-splines in each dimension to form the set $\mathcal{G}_0$. Our model selection is across combinations of tensor dimension and number of PCs used. Again we evaluate performance using both BIC and nearest neighbor residual correlation criteria.

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.

[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.

[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.