Improved Spatial Dependence-Robust Inference via Pre-whitening

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

April 27, 2026

Abstract. This paper presents a method to improve existing spatial dependence-robust inference procedures for spatial data. Our method involves augmenting a regression specification with functions of locations that reduce spatial correlation in regression scores and serves as a pre-whitener, followed by spatial HAC inference. We provide simulation evidence that our method results in substantial improvements in terms of statistical coverage properties for confidence intervals.

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 modeling 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 scenarios 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] offer further improvements when applicable, but still require tuning 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

*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].

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 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. 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. [3]

There are several ways to generate sensible basis functions which implement spatial pre-whitening. We 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 one B-spline is a step function, and an order two B-spline is a piece-wise linear "triangle/chevron" when non-zero, order three is a piece-wise parabola when non-zero, 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 bounds for the departure of coverage probability of HAC confidence intervals to 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].

Our analysis in this paper is for data that are mixing. However, in applications where the data are best modeled as being "trend stationary", mixing after de-trending, our method is easily applied to de-trended data. This includes examples in which functions of latitude and longitude have been projected out of observed spatial variables. In practice, if the researcher does not care about identifying spatial trends themselves, it will not be important to distinguish between spatial basis terms picking up trend variation

[3] There is no requirement that locations/distances be physical or geographic they can be determined by economic considerations.

versus picking up "transitory" variation. See Conley and Kelly (2025) for many applications of our method in circumstances both with and without spatial trends.

A related contribution is [Gonçalves and Ng, 2024] who are concerned with scenarios where predictions can be improved by augmenting estimates of a mean outcome with estimates of its associated idiosyncratic noise. This is feasible in applications where there is correlation across observations' idiosyncratic noises, e.g., because they follow a time series process. Our spatial basis terms on the other hand are functions of location, not other observations' prediction errors. We do not model and exploit a dependence structure between error terms.

Another related contribution is that in [Müller and Watson, 2024], who characterize a class of spatial unit root processes indexed on subsets of a Euclidean plane and 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. Section 7 concludes.

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 will assume below that the data are weakly dependent and that observations i and j approach independence as d_{ij} grows large.

We 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_0 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.

Our inference problem is to construct an interval estimate via a $1 - \alpha$ level confidence set $\widehat{C}$ for β_0, that satisfies

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

where ϵ 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 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. 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 $g : S \to [0, 1]$. 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 guidlines for choosing $\mathcal{G}$ will consider several candidate spline bases, e.g., $\mathcal{G}_1, \mathcal{G}_2$...,, in which for instance the knot points might differ.

To construct a confidence interval for a component of β_0, first estimate $[\widehat{\beta}, \widehat{\gamma}]$ with an OLS regression of Y_i on $[X_i, G_i]$. Then construct an estimate of the variance of $\widehat{\beta}$ using spatial HAC estimation with bandwidth $h > 0$ and kernel function k for the above regression. Let $\widehat{V}$ be the corresponding variance estimate.

For a component of β_0 that is of interest, $[\beta_0]_j$, let q_a be the ath quantile of the standard Gaussian distribution. random variable (i.e., of $\mathrm{N}(0, 1)$). We use the typical confidence interval estimator, $\widehat{C}_j$, that adds and subtracts a critical value, q_a, times the standard error estimate:

$$\widehat{C}_j = \big[\widehat{\beta}_j \pm q_{1-\alpha/2}[\widehat{V}]_{jj}^{1/2}\big].$$

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.

In 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 a frame. A frame is a tuple

$$F = (L_{\mathrm{mom}}, L_{\mathrm{mix}}, L_{\mathrm{cond}}, L_{\mathrm{growth}}, L_{\mathrm{basis}}, L_{\mathrm{kern}})$$

of positive constants satisfying $1 \leq \min(F)$. Each of the elements of F measures of regularity of the data and functions used in estimation, and can be thought of as finite upper bounds. They restrict moments, rank, mixing, metric regularity, the kernel, and pre-whitening basis, $\mathcal{G}$. We demonstrate properties of $\widehat{C}$ defined above relative to a given frame.

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

1. (*Linearity.*) $Y_i = X_i' \beta_0 + \varepsilon_i$ with $\mathrm{E}[\varepsilon_i X_j] = 0$ for $(i,j) \in S^2$.
2. (*Moments.*) $\mathrm{E}[\|X_i\|_{\text{Euclid}}^4] \leq L_{\text{mom}}$ as well as $\mathrm{E}[\varepsilon_i^4 | \mathcal{E}] \leq L_{\text{mom}}$ for $i \in S$ and non-null events $\mathcal{E}$ depending on (i.e., measurable with respect to the σ-algebra generated by) $\{X_j\}_{j \in S}$.
3. (*Mixing.*) For Z_A, Z_B random variables 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 depending on two arguments, $|\mathrm{E}[v(Z_A, Z_B) - v(Z_A, Z_B')]| \leq \exp(-d_{AB}/L_{\text{mix}}) \mathrm{E}[|v(Z_A, Z_B)| + |v(Z_A, Z_B')|]$, whenever the expectations are defined.
4. (*Conditioning.*) $\lambda_{\min}(|R|^{-1} \mathrm{E}[\left(\sum_{i,j \in R} a_{ij} \varepsilon_i X_j\right) \left(\sum_{i,j \in R} a_{ij} \varepsilon_i X_j\right)']) \geq 1/(L_{\text{cond}}|R|) \sum_{i,j \in R} a_{ij}^2$ and $\lambda_{\min}(|R|^{-1} \mathrm{E}[\left(\sum_{i \in R} c_i X_i\right) \left(\sum_{i \in R} c_i X_i\right)']) \geq 1/(L_{\text{cond}}|R|) \sum_{i \in R} c_i^2$ for all nonempty $R \subseteq S$, and constants c_i, a_{ij} with $i, j \in R$. Here $\lambda_{\min}$ denotes minimum eigenvalue and $|\cdot|$ denotes cardinality when applied to sets.
5. (*Metric Regularity.*) $d_{ij} \geq 1$ for $i \neq j \in S$ and $|B_{2r}(i)| \leq L_{\text{growth}} |B_r(i)|$ for $i \in S, r > 0$ where by convention, $B_r(i)$ is the closed 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 $\mathcal{A} : S \mapsto \mathcal{G}$ which assigns to every finite metric space S a collection functions $\mathcal{G} = \mathcal{A}(S)$, 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 least squares regression g on $\mathcal{G} \setminus \{g\}$. There are two constants associated to $\mathcal{G}$, which are $b \geq 1$ and $z \geq 1$, which measure the features related to the size of the supports of functions related to $g \in \mathcal{G}$. Define max-vol(z) to be the largest possible cardinality of a set of diameter z in S allowed by the growth and separation regularity conditions above. Estimation parameters in $\mathcal{T}_F$ satisfy the following conditions.

6. (*Kernel Regularity.*) $k(0) = 1$, $k(x) \in [0,1]$ for $x \in (0,1)$ and $k(x) = 0$ for $x \geq 1$. k Lipshitz continuous in that $|k(x) - k(x')| \leq L_{\text{kern}} |x - x'|$ for $x, x' \geq 0$.
7. (*Basis Regularity.*) For $g \in \mathcal{G}$, $b \leq |\{i \in S : |\tilde{g}(i)| \geq 1/L_{\text{basis}}\}|$ and $\mathrm{diam}(\mathrm{supp}(g)) \leq z$. $b \geq$ max-vol$(z)/L_{\text{basis}}$. For $i \in S$, $|\{g \in \mathcal{G} : B_z(i) \cap \mathrm{supp}(g) \neq \varnothing\}| \leq L_{\text{basis}}$. For $i \in S$ and $r > 0$, $|\{g \in \mathcal{G} : B_{2r}(i) \cap \mathrm{supp}(g) \neq \varnothing\}| \leq L_{\text{growth}} |\{g \in \mathcal{G} : B_r(i) \cap \mathrm{supp}(g) \neq \varnothing\}|$. For $i \in S$ and $g \in \mathcal{G}$, $|g(i)| \leq 1$ and $|\tilde{g}(i)| \leq L_{\text{basis}} \exp(-d_{i\mathrm{supp}(g)}/L_{\text{basis}})$. Here, $\mathrm{supp}(g) = \{i \in S : g(i) \neq 0\}$.

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.[4] 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 imposes standard regularity on the kernel function and bandwidth. A key part of Condition 6 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 several quantities related to the regularity of the spatial pre-whitening basis. This includes a recorded bound on the number of g supporting any i. For instance, in Figure 1, this number is 2. Tensor products of B-splines on lattices inherit bounds from their component B-splines. The two parameters b, z, which are treated on the same level as the kernel tuning parameter h, also control the regularity of $\mathcal{G}$. Specifically, b is a conditioning number, which measures the extend to which there is leftover variation in a basis term after projecting away other basis terms. The parameter z measures the maximum size of supports of basis terms g, and thus measures how localized the prewhitening basis. Note that $b/\text{max-vol}(z)$ measures the mass of remaining essential support points of $\tilde{g}$ of relative to the potential volume of the support of g, and is thus a measure of relative density. We note that it is sufficient that a collection of linear combination of elements $\mathcal{G}$ satisfy the conditions of $\mathcal{T}_F$ for the bounds of Theorem 1 below to hold.

To state the theorem, additionally define the doubling dimension of S as $\dim(S) = \log_2(L_{\text{growth}})$. Note that the existence of finite L_{growth} implies that $\text{max-vol}(z)$ grows polynomially in z.

Theorem 1. For every $\epsilon > 0$ and frame F, there is a threshold $M(\epsilon, F)$ depending only on ϵ and F such that if $n,\ z/\log(n)^4,\ n/(z\,\text{max-vol}(z)\text{max-vol}(h)),\ h/(z^2\,\text{max-vol}(z)),\ n/\text{max-vol}(z)^{8\dim(S)} > M(\epsilon, F)$, i.e., are sufficiently large, then for every data generating process in $\mathcal{P}_F$ and tuning parameters in $\mathcal{T}_F$,

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

One can equivalently restate the theorem as for every F, $|\Pr(\beta_0 \in \widehat{C}) - (1-\alpha)| < f_F(n, h, z, b)$ for a specific nonnegative-valued function f_F. We note that f_F is explicitly constructed in the proof of Theorem 1. f_F has the additional asymptotic vanishing property that for each F, $\lim f_F(n, h, z, b) = 0$ with the limit being over (n, h, z, b) as delineated by the statement of Theorem 1.

Comparing to the prior literature, the usual spatial HAC as in [Conley, 1996] also achieves asymptotically $1 - \alpha$ coverage, again up to a vanishing remainder term. Nevertheless, a good choice of $\mathcal{G}$ could simultaneously improve coverage probability and reduce the length of the confidence interval relative to other HAC variants. Additionally, there are potentially several sensible choices for pre-whitening bases $\mathcal{G}$. Choosing $\mathcal{G}$ is discussed in Section 5.

Because Theorem 1 develops finite sample bounds, all intermediate lemmas properties of also involve finite sample bounds. For example, 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].

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

We prove a Berry-Esseen-type central limit bound for spatial processes on arbitrary finite metric spaces. This is possible due to a particular decomposition of metric spaces called a padded partition, given by results in [Mendel and Naor, 2006], and an iterative Lindeberg swap procedure together with the replacement-type mixing as specified in Condition 3 above.

We comment here that max-vol(t) grows polynomially in t with polynomial depending only on L_{growth} (see Lemma 1). That is, there is an F-dependent polynomial q such that $n/\text{max-vol}(h) \to \infty$ is equivalent to $n/q(h) \to \infty$, for instance. Sequences (n, h, z, b) satisfying the above exist: it suffices to set z larger than $\log^4(n)$, b on the order of max-vol(z), and h smaller than $n^{1/a}$ for an a depending only on F.

There are several technical hurdles that our analysis needs to handle which arise from the fact that there are essentially two interacting non-parametric-type estimation procedures being used at the same time. In particular, the number of terms in the the pre-whitening basis does not enter into the bounds for Theorem 1 and there is no finite bound on the bandwidth parameter. Additionally, the pre-whitening regressors are non-stationary, and though their support is assumed localized, their residuals when regressed on each other may have a support all of S.

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 β_0, that is, consider $p = 1$. The case $p > 1$ is analogous.[5]

The following conventions are used. A collection of random variables Z_i at $i \in S$ are said to mix with delay ν and rate L_{mix} if for any measurable function v and random variables Z_A, Z_B, Z'_B depending (measurably) on $\{Z_i\}_{i\in A}, \{Z_i\}_{i\in B}$, and Z'_B an independent of copy of Z_B, it holds that $|\mathrm{E}[v(Z_A, Z_B) - v(Z_A, Z'_B)]| \leq \exp(-(d_{AB} - \nu)/L_{\text{mix}})(\mathrm{E}[|v(Z_A, Z_B)|] + \mathrm{E}[|v(Z_A, Z'_B)|])$ whenever the expectations exist. $B_x(i) = \{j \in S : d_{ij} \leq x\}$ denotes the closed ball around i. All log operations are base 2.

Lemma 1 (Cardinalities of Closed Balls and Diagonals). For $i \in S$ and $x \geq 1$, $|B_x(i)| \leq L_{\text{growth}}^{\log x+2}$. For any $T \subseteq S$, $x \geq 1$, setting $\Delta = \{i \in T^2 : d_{i_1 i_2} \leq x\}$ gives $|\Delta| \leq |T| L_{\text{growth}}^{\log x+2}$.

[5] Note, the combination of $(L_{\text{mom}}, L_{\text{cond}})$ can be used to infer an upper bound on p, so $p > 1$ requires no additional information entered into F.

Proof of Lemma 1. For $i \in T$, $|B_x(i)| \leq L_{\text{growth}}|B_{x/2}(i)| \leq ... \leq L_{\text{growth}}^{\lceil \log x \rceil + 1}|B_{x/2^{\lceil \log x \rceil + 1}}(i)|$, where $\lceil \log x \rceil$ denotes least integer $\geq \log x$. Note that $x/2^{\lceil \log x \rceil + 1} < 1$ gives $|B_{x/2^{\lceil \log x \rceil + 1}}(i)| = |\{i\}| = 1$ and that $\lceil \log x \rceil + 1 \leq \log x + 2$. Note, this implies that for all $i \in T, x \geq 1$, $|B_x(i)| \leq L_{\text{growth}}^{\log x + 2}$, a fact which will be used repeatedly. Then $|\Delta| \leq \sum_{i \in T} |B_x(i) \cap T| \leq |T| L_{\text{growth}}^{\log x + 2}$. ■

Lemma 2 (Decomposition of Cartesian Products of Diagonals). Let $x, y \geq 1$. Let $T \subseteq S$. Define subsets of S^4:

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

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

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

Proof of Lemma 2. 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 $B_x(\pi i_1)$ and $B_x(\pi i_3)$ 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 $d_{\{\pi i_1, \pi i_2\}\{\pi i_3, \pi i_4\}} > x$. So $i \in C_3$. Next bound the cardinalities of A, C_1. Let $A^{1/2} = \{i \in T^2 : i_2 \in B_y(i_1)\}$. Then $|A^{1/2}| \leq |T| \max_{i \in S} |B_y(i)|$. As in Lemma 1, $|B_y(i)| \leq L_{\text{growth}}^{\log y + 2}$. Then $A = A^{1/2} \times A^{1/2}$ gives $|A| \leq |A^{1/2}|^2$. Next, $|C_1|$ is bounded analogously. Finally, by inclusion, $|C_2| + |C_3| \leq |A|$. ■

Lemma 3 (Padded Partitions of Metric Spaces). Let $x \geq 1$ and $y \geq 8x$. There is a collection $\mathcal{R}^\circ$ of disjoint subsets of S, together with a boundary $\partial\mathcal{R} = S \setminus \cup_{R \in \mathcal{R}^\circ} R$, i.e.,

$$S = \partial\mathcal{R} \cup \bigcup_{R \in \mathcal{R}^\circ} R,$$

with the following properties. Either $|\mathcal{R}^\circ| = 1$ or else all of the following hold: $R, R' \in \mathcal{R}^\circ$, $d_{RR'} \geq x$ and $L_{\text{growth}}^{\log(y)+2} \leq |R| \leq 3L_{\text{growth}}^{\log(y)+2}$, the boundary has $|\partial\mathcal{R}| \leq n(1 - L_{\text{growth}}^{-48x/y})$, and $|\mathcal{R}^\circ| \geq \frac{1}{3} n L_{\text{growth}}^{-48x/y - \log y - 2}$. Additionally, $n > 3L_{\text{growth}}^{48x/y + \log y + 2}$ is sufficient for $|\mathcal{R}^\circ| > 1$.

Proof of Lemma 3. Let $x > 0$ and let $y \geq 8x$. By Lemma 3.1 in [Mendel and Naor, 2006], there is a distribution Pr over partitions $\mathcal{R}$ of S such that for every $i \in S$, $\Pr(\text{diam}(\mathcal{R}(i)) \leq y) = 1$ and

$$\Pr(B_x(i) \subseteq \mathcal{R}(i)) \geq \left(\frac{|B_{y/8}(i)|}{|B_y(i)|} \right)^{16x/y}.$$

Note, this distribution is called a padded decomposition of S, and exists for all finite metric spaces. Let $\partial\mathcal{R} = \{i : \mathcal{R}(i) \neq \mathcal{R}(i') \text{ for some } i' \in B_x(i)\}$. Thus, the expected number of boundary points can by obtained by summing expected values of indicators of inclusion:

$$\mathrm{E}[|\partial\mathcal{R}|] = \sum_{i \in S} \Pr(B_x(i) \not\subseteq \mathcal{R}(i)).$$

Applying $\frac{|B_{y/8}(i)|}{|B_y(i)|} \geq L_{\text{growth}}^{-3}$ with the first bound gives $\mathrm{E}[|\partial\mathcal{R}|] \leq n(1 - L_{\text{growth}}^{-48x/y})$. By the pigeonhole principal, there exists at least one partition draw $\mathcal{R}$ with $\partial\mathcal{R}$ with that bound. Fix that partition. Note

that $\operatorname{diam}(R) \le y$ implies that $R \subseteq B_i(y)$ for any i realizing $\max\{d_{ij} : i, j \in R\}$. Thus $|R| \le |B_i(y)| \le L_{\text{growth}}^{\log(y)+2}$.

Let $\mathcal{R}^\circ = \{R \setminus \partial\mathcal{R} : R \in \mathcal{R}\}$. Refine $\mathcal{R}^\circ$ repeatedly as follows. If there are at least two $R, R' \in \mathcal{R}^\circ$ with $|R|, |R'| < L_{\text{growth}}^{\log(y)+2}$, then select two such arbitrarily, and update $\mathcal{R}^\circ \leftarrow (\mathcal{R}^\circ \setminus \{R, R'\}) \cup \{R \cup R'\}$. As n is finite, this process must terminate. At the end, all but at most one R have $|R| \in [L_{\text{growth}}^{\log(y)+2}, 2L_{\text{growth}}^{\log(y)+2}]$. If there are no leftover R with $|R| < L_{\text{growth}}^{\log(y)+2}$, then keep this $\mathcal{R}^\circ$, and the bound on the sizes of $|R|$ in the statement of the lemma thus holds. If there is still R with cardinality less than $L_{\text{growth}}^{\log(y)+2}$, then either $|\mathcal{R}^\circ| = 1$, or choose another R' from $\mathcal{R}^\circ$, which means $|R'| \in L_{\text{growth}}^{\log(y)+2}, 2L_{\text{growth}}^{\log(y)+2}]$ and note $|R \cup R'| \le 3L_{\text{growth}}^{\log(y)+2}$. Do $\mathcal{R}^\circ \leftarrow (\mathcal{R}^\circ \setminus \{R, R'\}) \cup \{R \cup R'\}$ a final time, ensuring $\mathcal{R}^\circ$ has all elements R satisfying $|R| \in [L_{\text{growth}}^{\log(y)+2}, 3L_{\text{growth}}^{\log(y)+2}]$. Then $S = \partial\mathcal{R} \cup \bigcup_{R \in \mathcal{R}^\circ} R$ is a disjoint union with $d_{RR'} \ge x$ for $R, R' \in \mathcal{R}^\circ$ and $|\mathcal{R}^\circ| \ge (n - |\partial\mathcal{R}|)\frac{1}{3}L_{\text{growth}}^{-\log(y)-2} \ge \frac{1}{3}nL_{\text{growth}}^{-48x/y-\log(y)-2}$. ■

The next lemmas are properties of random variables over S. For $T \subseteq S$ and $x \ge 1, \nu \ge 0$, define

$$f_{\text{LLN}}(T, x, \nu) = |T|^{-1}L_{\text{growth}}^{\log x+2} + 2\exp(-(x-\nu)/L_{\text{mix}}),$$
$$f_{\text{4TH}}(T, x, \nu) = 4!L_{\text{growth}}^{2\log 3x+4} + 2|T|^2\exp(-(x-\nu)/L_{\text{mix}}).$$

Lemma 4 (Law of Large Numbers). Let Z_i be random variables with finite second moments at all $i \in T \subseteq S$ that mix with delay $\nu \ge 0$. Let $c > 0, x \ge 1$. Then

$$\Pr\left(|T|^{-1}\left|\sum_{i\in T} Z_i - \mathrm{E}[Z_i]\right| > c\right) \le c^{-2}f_{\text{LLN}}(T, x, \nu)\max_{i\in T}\operatorname{var}(Z_i).$$

Proof of Lemma 4. By Lemma 1, $|\Delta| \le |T|L_{\text{growth}}^{\log x+2}$. Then, $\mathrm{E}[(|T|^{-1}\sum_{i\in T}(Z_i - \mathrm{E}[Z_i]))^2] = |T|^{-2}\mathrm{E}[\sum_{i\in\Delta}(Z_{i_1} - \mathrm{E}[Z_{i_1}])(Z_{i_2} - \mathrm{E}[Z_{i_2}]) + \sum_{i\in T^2\setminus\Delta}(Z_{i_1} - \mathrm{E}[Z_{i_1}])(Z_{i_2} - \mathrm{E}[Z_{i_2}]) \le |T|^{-2}\mathrm{E}[\sum_{i\in\Delta}|\operatorname{cov}(Z_{i_1}, Z_{i_2})| + \sum_{i\in T^2\setminus\Delta}|\operatorname{cov}(Z_{i_1}, Z_{i_2})|$ which by Cauchy-Schwarz inequality and the mixing condition of this lemma is $\le |T|^{-2}(|\Delta|\max_{i\in T}\operatorname{var}(Z_i) + |T|^2 2\exp(-(x-\nu)/L_{\text{mix}})\max_{i\in T}\operatorname{var}(Z_i)) \le |T|^{-2}\left(|T|L_{\text{growth}}^{\log x+2} + |T|^2 2\exp(-(x-\nu)/L_{\text{mix}})\right)\max_{i\in T}\operatorname{var}(Z_i)$. Simplifying gives $\le (|T|^{-1}L_{\text{growth}}^{\log x+2} + 2\exp(-(x-\nu)/L_{\text{mix}}))\max_{i\in T}\operatorname{var}(Z_i) = f_{\text{LLN}}(T, x, \nu)\max_{i\in T}\operatorname{var}(Z_i)$. Chebyshev's inequality gives the lemma. ■

Lemma 5 (General Fourth Moment Bounds). Let $T \subseteq S$ nonempty and let Z_i be mean 0 random variables at $i \in T$ that mix with delay $\nu \ge 0$. Let $Z_T = |T|^{-1/2}\sum_{i\in T} Z_i$. Let $x \ge 1$. Then

$$\mathrm{E}[Z_T^4] \le f_{\text{4TH}}(T, x, \nu)\max_{i\in T}\mathrm{E}[Z_i^4].$$

Proof of Lemma 5. Let $x \ge 1$ and let $A^\circ = \{i \in T^4 : \text{ no permutation of } i \text{ is in } A\ \}$ where A is the set defined in Lemma 2 using $y = 3x$. If $i \in A^\circ$ then there is a permutation of i such that $d_{\pi i_1, \{\pi i_2, \pi i_3, \pi i_4\}} > x$. To see this, draw a square with corners labeled i_1, i_2, i_3, i_4. As $i \in A^\circ$, at least one horizontal side has corresponding distance $> y = 3x$, and similarly one vertical side has distance $> 3x$. There is thus a corner, j, incident to both a horizontal and vertical side with distance $> x$. Let a, b be the clockwise and counterclockwise neighboring corners of j and o be the corner opposite j. If $d_{jo} > x$, then $d_{j(\{o,a,b\}} > x$. Else, if $d_{jo} < x$ then $d_{ab} > 3x$ as otherwise a permutation of i is in A. Then by triangle inequality, $d_{ao} \ge d_{ja} - d_{jo} > 2x$, and so in this case $d_{a\{j,b,o\}} > x$. If the standalone element is i_1, say, then $|\mathrm{E}[Z_{i_1}Z_{i_2}Z_{i_3}Z_{i_4}]| = |\mathrm{E}[Z_{i_1}Z_{i_2}Z_{i_3}Z_{i_4} - 0] = |\mathrm{E}[Z_{i_1}Z_{i_2}Z_{i_3}Z_{i_4} - Z'_{i_1}Z_{i_2}Z_{i_3}Z_{i_4}] \le (\mathrm{E}[|Z_{i_1}Z_{i_2}Z_{i_3}Z_{i_4}|] +$

$\mathrm{E}[|Z'_{i_1} Z_{i_2} Z_{i_3} Z_{i_4}|]) \exp(-(x-\nu)/L_{\mathrm{mix}}) \le 2 \max_{j \in T} \mathrm{E}[Z_j^4] \exp(-(x-\nu)/L_{\mathrm{mix}})$ by the lemma's mixing condition followed by Hölder's inequality, where Z'_{i_1} is an independent of $Z_{i_2}, Z_{i_3}, Z_{i_4}$ copy of Z_{i_1}. Then it follows that $\max_{i \in A^\circ} |\mathrm{E}[Z_{i_1} Z_{i_2} Z_{i_3} Z_{i_4}]| \le 2 \max_{j \in T} \mathrm{E}[Z_j^4] \exp(-(x-\nu)/L_{\mathrm{mix}})$. Using Lemma 2, $|T^4 \setminus A^\circ| \le 4!|A| \le 4!|T|^2 L_{\mathrm{growth}}^{2 \log 3x+4}$ after which it follows that $\mathrm{E}[Z_T^4] = |T|^{-2}(\sum_{i \in A^\circ} + \sum_{i \notin A^\circ}) \mathrm{E}[Z_{i_1} Z_{i_2} Z_{i_3} Z_{i_4}]$ and this is further bounded by $\le \left(|T|^{-2}|A^\circ| 2 \exp(-(x-\nu)/L_{\mathrm{mix}}) + |T|^{-2} 4! |T|^2 L_{\mathrm{growth}}^{2 \log 3x+4}\right) \max_{j \in T} \mathrm{E}[Z_j^4]$. Simplifying and applying $|T|^{-2}|A^\circ| \le |T|^2$ (which holds as $A^\circ \subseteq T^4$ so $|A^\circ| \le |T|^4$) gives the proof. ■

Let

$$\begin{aligned} f_{\mathrm{CLT}}(n, \rho, x, \nu) = {} & 15.12 n^{-1/2} \rho L_{\mathrm{growth}}^{3/x + \log(x) + 5/2} \\ & + 6n \exp(-(x-\nu)/L_{\mathrm{mix}}) \\ & + 16 \rho^{4/9} (x^{-1/3} + n^{2/3} \exp(-(x-\nu)/(3L_{\mathrm{mix}}))) L_{\mathrm{growth}}^{2/x + 2/3}. \end{aligned}$$

Lemma 6 (Central Limit Theorem). Let Z_i be mean zero random variables at $i \in S$ that mix with delay ν. Let $\Xi = \sum_{i \in S} Z_i$. Let $\sigma^2 = \mathrm{var}(\Xi)$. Let $t \in \mathbb{R}$. Let $x \ge 1$. Suppose

$$\rho \ge \frac{\max_{T \subseteq S} \mathrm{E}[||T|^{-1/2} \sum_{i \in T} Z_i|^3]}{\min_{\varnothing \ne T \subseteq S} \mathrm{E}[(|T|^{-1/2} \sum_{i \in T} Z_i)^2]^{3/2}}.$$

Then

$$|\Pr(\sigma^{-1} \Xi \le t) - \Phi(t)| \le f_{\mathrm{CLT}}(n, \rho, x, \nu).$$

Proof of Lemma 6. Let $\mathcal{R}^\circ$ be the partition from Lemma 3 with $x \ge 1$ and $y = 8x^2$. Let $Z_R = |R|^{-1/2} \sum_{i \in R} Z_i$. Equate $\Xi = \sum_{R \in \mathcal{R}^\circ} |R|^{1/2} Z_R + r$ with remainder $r = |\partial \mathcal{R}|^{1/2} Z_{\partial \mathcal{R}}$. Let Z'_R be independent copies of Z_R. Note that $|\mathcal{R}^\circ| \ge n \frac{1}{3} L_{\mathrm{growth}}^{-48x/(8x^2) - \log(8x^2) - 2}$ which simplifies to and results in

$$|\mathcal{R}^\circ| \ge n \frac{1}{3} L_{\mathrm{growth}}{}^{-6/x - 2\log(x) - 5} \text{ and } |\partial \mathcal{R}| \le n(1 - L_{\mathrm{growth}}^{-6/x}) \text{ and } \min_{R \in \mathcal{R}^\circ} |R| \ge L_{\mathrm{growth}}^{5 + 2\log(x)}.$$

Let $m = |\mathcal{R}^\circ|$. Order $R \in \mathcal{R}^\circ$ arbitrarily with $R_1, ..., R_m$. Then let $\Xi_0 = \Xi - r$ and $\Xi_l = \Xi_{l-1} - |R_l|^{1/2} Z_{R_l} + |R_l|^{1/2} Z'_{R_l}$. Let σ_l be the variance of Ξ_l. Then Ξ_m, being a sum of independent random variables, by the Berry-Esseen central limit theorem, satisfies $|\Pr(\sigma_m^{-1} \Xi_m \le t) - \Pr(N(0,1) \le t)| \le 0.56 \sum_{R \in \mathcal{R}^\circ} \mathrm{E}[|R|^{3/2} |Z_R|^3] / \left(\sum_{R \in \mathcal{R}^\circ} \mathrm{E}[|R|^1 Z_R^2]\right)^{3/2}$. This can be further simplified by the bound $\le 0.56 m^{-1/2} \max_{R \in \mathcal{R}^\circ} |R|^{3/2} \max_{R \in \mathcal{R}^\circ} \mathrm{E}[|Z_R|^3] / (\min_{R \in \mathcal{R}^\circ} |R|^{3/2} \min_{R \in \mathcal{R}^\circ} \mathrm{E}[Z_R^2]^{3/2})$. Note by Lemma 3 that $\max_{R \in \mathcal{R}^\circ} |R|^{3/2} / \min_{R \in \mathcal{R}^\circ} |R|^{3/2} \le 3^{3/2}$. Using that $0.56 \times 3^{3/2} \times \frac{1}{3}^{-1/2} = 5.04$, as well as the definition of ρ gives

$$|\Pr(\sigma_m^{-1} \Xi_m \le t) - \Pr(N(0,1) \le t)| \le 5.04 n^{-1/2} L_{\mathrm{growth}}^{3/x + \log(x) + 5/2} \rho.$$

As for every $l \le m$, $\Pr(\Xi_l \le \sigma_m t) - \Pr(\Xi_{l-1} \le \sigma_m t)| \le 2 \exp(-(x-\nu)/L_{\mathrm{mix}})$, summing, and using $m \le n$ gives

$$|\Pr(\sigma_m^{-1} \Xi_0 \le t) - \Pr(\sigma_m^{-1} \Xi_m \le t)| \le 2n \exp(-(x-\nu)/L_{\mathrm{mix}}).$$

Let ϵ_0 be the sum of the two above bounded quantities. Let $t_* > 0$ be a threshold. For $|t| \ge t_*$, by union bound and Chebyshev's inequality, $\sup_{|t| \ge t_*} |\Pr(\Xi/\sigma \le t) - \Pr(\Xi_0/\sigma_m \le t)| \le \Pr(|\Xi| \ge \sigma t_*) + \Pr(|\Xi_0| \ge$

$\sigma_m t_*) \le 1/t_*^2 + (\sigma_0/\sigma_m)^2/t_*^2$. Note that $\Xi = \Xi_0 + r$, so for any $u > 0$, $\Pr(\Xi \le \sigma t) - \Pr(\Xi_0 \le \sigma t) \le \Pr(|r| > u) + \sup_{a\in\mathbb{R}} \Pr(a < \Xi_0 \le a+u)$. Then, $|\Pr(\Xi \le \sigma t) - \Pr(\Xi_0 \le \sigma_m t)| \le \Pr(|r| > u) + \sup_{a\in\mathbb{R}} \Pr(a < \Xi_0 \le a+u+|\sigma-\sigma_m||t|)$. Applying the above bound with $|t| < t_*$, the two bounds can be added together to get $|\Pr(\sigma^{-1}\Xi \le t) - \Pr(\sigma_m^{-1}\Xi_0 \le t)| \le (1+(\sigma_0/\sigma_m)^2)/t_*^2 + \mathrm{var}(r)/u^2 + \sup_{a\in\mathbb{R}} \Pr(a < \Xi_0 \le a+u+|\sigma-\sigma_m|t_*)$. Comparing to $\Phi(t)$ via ϵ_0 given $\sup_{t\in\mathbb{R}} |\Pr(\sigma^{-1}\Xi \le t) - \Pr(\sigma_m^{-1}\Xi_0 \le t)| \le (1+(\sigma_0/\sigma_m)^2)/t_*^2 + \mathrm{var}(r)/u^2 + (u+|\sigma-\sigma_m|t_*)/(\sigma_m\sqrt{2\pi}) + 2\epsilon_0$. Let $u = (\mathrm{var}(r)\sigma_m\sqrt{2\pi})^{1/3}$ and $t_* = ((1+(\sigma_0/\sigma_m)^2)\sqrt{2\pi}\sigma_m/|\sigma-\sigma_m|)^{1/3}$. This gives

$$\sup_{t\in\mathbb{R}} |\Pr(\sigma^{-1}\Xi \le t) - \Pr(\sigma_m^{-1}\Xi_0 \le t)| \le 2\left(\left(\frac{\mathrm{var}(r)}{2\pi\sigma_m^2}\right)^{1/3} + \left(\left(1+\frac{\sigma_0^2}{\sigma_m^2}\right)\frac{(\sigma-\sigma_m)^2}{2\pi\sigma_m^2}\right)^{1/3}\right) + 2\epsilon_0.$$

Note that $\sigma_m^2 \ge m \min_{R\in\mathcal{R}^\circ} |R| \min_{R\in\mathcal{R}^\circ} \mathrm{E}[Z_R^2]$. Note that $\mathrm{var}(r) \le |\partial\mathcal{R}|\mathrm{E}[Z_{\partial\mathcal{R}}^2]$. Note that $|\sigma - \sigma_0| \le \mathrm{var}(r)^{1/2}$ by Minkowski's inequality bounding together the standard deviations of mean zero random variables $\Xi_0 + r$ and Ξ_0. Note that $|\sigma_{l-1} - \sigma_l| \le (\sigma_{l-1} + \sigma_l)\exp(-(x-\nu)/L_{\mathrm{mix}})$ after applying the assumed mixing condition. Note that $\sigma_l \le n^{1/2} \max_{T\subseteq S} \mathrm{E}[Z_T^2]^{1/2}$. Therefore summing leaves $|\sigma_0 - \sigma_m| \le 2mn^{1/2} \max_{T\subseteq S} \mathrm{E}[Z_T^2]^{1/2} \exp(-(x-\nu)/L_{\mathrm{mix}})$. And $(\sigma-\sigma_m)^2 = (\sigma - \sigma_0 + \sigma_0 - \sigma_m)^2 \le (2|\partial\mathcal{R}| + 4m^2 n \exp(-(x-\nu)/L_{\mathrm{mix}})) \max_{T\subseteq S} \mathrm{E}[Z_T^2]$. Note that $1 + \sigma_0^2/\sigma_m^2 \le \frac{(n-|\partial\mathcal{R}|)}{(n-|\partial\mathcal{R}|)/3} \frac{\max_{T\subseteq S} \mathrm{E}[Z_T^2]}{\min_{T\subseteq S} \mathrm{E}[Z_T^2]} \le 1 + 3\frac{\max_{T\subseteq S} \mathrm{E}[Z_T^2]}{\min_{T\subseteq S} \mathrm{E}[Z_T^2]}$. Putting the above together, yields the following bound.

$$\frac{2}{(2\pi)^{1/3}} \frac{|\partial\mathcal{R}|^{1/3} + \left(2|\partial\mathcal{R}| + 4m^2 n \exp(-(x-\nu)/L_{\mathrm{mix}})\right)^{1/3} \left(1 + 3\frac{\max_{T\subseteq S} \mathrm{E}[Z_T^2]}{\min_{T\subseteq S} \mathrm{E}[Z_T^2]}\right)^{1/3}}{(m \min_{R\in\mathcal{R}^\circ} |R|)^{1/3}} \left(\frac{\max_{T\subseteq S} \mathrm{E}[Z_T^2]}{\min_{T\subseteq S} \mathrm{E}[Z_T^2]}\right)^{1/3}.$$

Note that $(\mathrm{E}[Z_T^2]/\mathrm{E}[Z_{T'}^2])^{1/3} \le (\mathrm{E}[|Z_T|^3]^{2/3}/\mathrm{E}[Z_{T'}^2])^{1/3} = (\mathrm{E}[|Z_T|^3]/\mathrm{E}[Z_{T'}^2]^{3/2})^{1/3\times 2/3} \le \rho^{2/9}$. For scalars $a, b, c \ge 0$ note $ca^{1/3} + (2a+2b)^{1/3} \le (c + 2^{1/3})(a^{1/3} + b^{1/3})$. Applying above bounds surrounding $\mathcal{R}^\circ$, this yields the reduction $\frac{2(1/2\pi+2)^{1/3}}{(1/3)^{1/3}} \frac{(1-L_{\mathrm{growth}}^{-6/x})^{1/3} + n^{2/3}\exp(-(x-\nu)/3L_{\mathrm{mix}}))(1+\rho^{2/3})}{L_{\mathrm{growth}}^{(-6/x-2\log(x)-5+\log(8x^2))/3}} \rho^{2/9} + 2\epsilon_0$. For the display in the statement of the theorem, the following simplifications are used. Note that $2(1/2\pi+2)^{1/3}(1/3)^{-1/3} < 4$. Note also that $(1-L_{\mathrm{growth}}^{-6/x})^{1/3} \le (6L_{\mathrm{growth}})^{1/3} x^{-1/3} \le 2x^{-1/3} L_{\mathrm{growth}}^{1/3}$, using that $6^{1/3} \le 2$. This factor of 2 can be absorbed into the leading constant. Note that exponent in the denominator simplifies to $-2/x - 2/3$. This leaves $8(x^{-1/3} + n^{2/3}\exp(-(x-\nu)/2L_{\mathrm{mix}}))L_{\mathrm{growth}}^{2/x+2/3}$. Finally, all of the ρ-related terms can be gathered and bounded by $(1+3\rho^{2/3})^{1/3}\rho^{2/9} \le 2\rho^{4/9}$. The leading constant for this term is then $8 \times 2 = 16$. ■

Next are properties of $\widehat{\xi}, \widehat{\eta}$ and $\widehat{\zeta}$, which are defined as linear least squares regression coefficients X_i, ε_i and Y_i on G_i. Denote

$$\tilde{X}_i = X_i - G_i'\widehat{\xi},\ \tilde{\varepsilon}_i = \varepsilon_i - G_i'\widehat{\eta},\ \tilde{Y}_i = Y_i - G_i'\widehat{\zeta}.$$

For $r > 0$ and $g \in \mathcal{G}$, let $\mathcal{K}_g(r) = \{i \in S : d_{i\mathrm{supp}(g)} \le r\}$. Define

$$\widehat{\xi}_g^{\sharp(r)} = \Big(\sum_{i\in S} \tilde{g}(i)^2\Big)^{-1} \sum_{i\in\mathcal{K}_g(r)} X_i\tilde{g}(i), \qquad \widehat{\xi}_g^{\flat(r)} = \Big(\sum_{i\in S} \tilde{g}(i)^2\Big)^{-1} \sum_{i\in S\setminus\mathcal{K}_g(r)} X_i\tilde{g}(i),$$
$$X_i^{\sharp(r)} = X_i - \sum_{g\in\mathcal{G}} g(i)\widehat{\xi}_g^{\sharp(r)}, \qquad X_i^{\flat(r)} = -\sum_{g\in\mathcal{G}} g(i)\widehat{\xi}_g^{\flat(r)},$$

It is helpful to note that $\mathrm{E}[\tilde{X}_j\varepsilon_i] = \mathrm{E}[X_j^{\sharp(r)}\varepsilon_i] = \mathrm{E}[X_j^{\flat(r)}\varepsilon_i]$ as each of $\tilde{X}_j$,$X_j^{\sharp(r)}$, $X_j^{\flat(r)}$ are linear combinations of $X_{j'}, j' \in S$.

Lemma 7 (Localization of Basis Projections). Let $r > 0$. There are decompositions,

$$\tilde{X}_i = X_i^{\sharp(r)} + X_i^{\flat(r)},\ \widehat{\xi}_g = \widehat{\xi}_g^{\sharp(r)} + \widehat{\xi}_g^{\flat(r)}.$$

Proof of Lemma 7. By the Frisch-Waugh theorem, $\widehat{\xi}_g = (\sum_{i\in S} \tilde{g}(i)^2)^{-1} \sum_{i\in S} X_i \tilde{g}(i) = \widehat{\xi}_g^{\sharp(r)} + \widehat{\xi}_g^{\flat(r)}$ and thus $\tilde{X}_i = X_i^{\sharp(r)} + X_i^{\flat(r)}$. Note that the analogous statement holds for Y_i and ε_i. ■

Let

$$f_{\mathrm{TRNC}}(r,x) = b^{-4} L_{\mathrm{basis}}^{12} L_{\mathrm{mom}} n^2 \exp(-4r/L_{\mathrm{basis}}) f_{4\mathrm{TH}}(S,x,0).$$

Lemma 8 (Approximate Preservation of Mixing). Let $r > 0$ and $x \geq 1$. Then $X_i^{\sharp(r)}$ depends only on $\{X_{i'}\}_{i'\in B_{2r+2z}(i)}$ and $\widehat{\xi}_g^{\sharp(r)}$ depends only on $\{X_i\}_{i\in\mathcal{K}_g(r)}$. Furthermore,

$$\mathrm{E}[(\widehat{\xi}_g^{\flat(r)})^4] \leq f_{\mathrm{TRNC}}(r,x) \ \text{ and } \ \mathrm{E}[(X_i^{\flat(r)})^4] \leq L_{\mathrm{basis}}^4 f_{\mathrm{TRNC}}(r,x).$$

Additionally, $\{(Y_i^{\sharp(r)}, X_i^{\sharp(r)})\}_{i\in S}$ mix with delay $4z + 4r$.

Proof of Lemma 8. For the denominator, by Regularity Condition 7, $\sum_{i\in S} \tilde{g}(i)^2 \geq \sum_{i:|\tilde{g}(i)|\geq 1/L_{\mathrm{basis}}} 1/L_{\mathrm{basis}}^2$ $\geq |\{i : |\tilde{g}(i)| \geq 1/L_{\mathrm{basis}}\}|/L_{\mathrm{basis}}^2 \geq b/L_{\mathrm{basis}}^2$. Therefore, $\left(\sum_{i\in S} \tilde{g}(i)^2\right)^{-4} \leq b^{-4} L_{\mathrm{basis}}^8$. For the numerator, use Apply Lemma 5 to yield that $(|S \setminus \mathcal{K}_g(r)|^{1/2})^4 \mathrm{E}[(\frac{1}{|S\setminus\mathcal{K}_g(r)|^{1/2}} \sum_{i\in S\setminus\mathcal{K}_g(r)} X_i \tilde{g}(i))^4] \leq |S|^2 f_{4\mathrm{TH}}(S \setminus \mathcal{K}_g(r), x, 0) L_{\mathrm{basis}}^4 L_{\mathrm{basis}}^4 \exp(-4r/L_{\mathrm{basis}})$. Combining the numerator and denominator gives $\mathrm{E}[(\xi_g^{\flat(r)})^4] \leq b^{-4} L_{\mathrm{basis}}^{12} n^2 \exp(-4r/L_{\mathrm{basis}}) f_{4\mathrm{TH}}(S,x,0) \max_{i\in S} \mathrm{E}[X_i^4]$, and subsequently the Lemma's first stated bound. Then the bound on $\mathrm{E}[(X_i^{\flat(r)})^4]$ follows from the cardinality bound on $|\mathcal{G}_i| \leq L_{\mathrm{basis}}$ from Regularity Condition 7, where $\mathcal{G}_i$ is defined as $\mathcal{G}_i = \{g \in \mathcal{G} : i \in \mathrm{supp}(g)\}$. With $|g(i)| \leq 1$, $\mathrm{E}[(X_i^{\flat(r)})^4] \leq |\mathcal{G}_i|^3 \sum_{g\in\mathcal{G}_i} g(i)^4 \mathrm{E}[(\widehat{\xi}_g^{\flat(r)})^4] \leq L_{\mathrm{basis}}^4 \max_{g\in\mathcal{G}} \mathrm{E}[(\widehat{\xi}_g^{\flat(r)})^4]$. Finally, note that $\mathrm{diam}(\cup_{g\in\mathcal{G}_i} \mathcal{K}_g(r)) \leq 2z + 2r$ because all $K_g(r)$ comprising this union have a common point of support, namely i. In fact, then $X_i^{\sharp(r)}$ depends only on $X_{i'}$ from $i' \in B_{2z+2r}(i)$. Then to show mixing, suppose that Z_A and Z_B depend on $\{(Y_i^{\sharp(r)}, X_i^{\sharp(r)})\}_{i\in A}$ and $\{(Y_i^{\sharp(r)}, X_i^{\sharp(r)})\}_{i\in B}$. Let $A^{\sharp(r)} = \cup_{i\in A} B_{2r+2z}(i), B^{\sharp(r)} = \cup_{i\in B} B_{2r+2z}(i)$. As $d_{A^{\sharp(r)}B^{\sharp(r)}} \geq d_{AB} - (4r + 4z)$. Apply regularity Condition 3 to the original $\{(Y_i, X_i)\}$ to get that $\{(Y_i^{\sharp(r)}, X_i^{\sharp(r)})\}_{i\in S}$ mix with delay $4r + 4z$. Analogous holds for $\widehat{\eta}_g$ and $\widehat{\zeta}_g$ and $\tilde{\varepsilon}_i$ and $\tilde{Y}_i$. ■

Let

$$f_{\mathrm{SPL}}(r,x) = 8b^{-4} L_{\mathrm{basis}}^{12} L_{\mathrm{mom}} \Big(L_{\mathrm{growth}}^{2\log(2r+2z)+4} \max_{g\in\mathcal{G}} f_{4\mathrm{TH}}(\mathcal{K}_g(r), x, 0) + n^2 \exp(-4r/L_{\mathrm{basis}}) f_{4\mathrm{TH}}(S,x,0) \Big),$$

$$f_{\mathrm{PAR}}(r,x) = 8L_{\mathrm{mom}} + 8L_{\mathrm{basis}}^4 f_{\mathrm{SPL}}(r,x),$$

$$f_{\mathrm{COND}} = \max(0, 1 - L_{\mathrm{basis}}^4/b)^2 / L_{\mathrm{cond}}.$$

Note that $f_{\mathrm{SPL}}(r,x)$ after expanding, is bounded by $f_{\mathrm{SPL}}(r,x) \leq 8b^{-4} L_{\mathrm{basis}}^{12} L_{\mathrm{mom}} (24 L_{\mathrm{growth}}^{2\log(2r+2z)+2\log(3x)+8} + 2L_{\mathrm{growth}}^{4\log(2r+2z)+8} e^{-x/L_{\mathrm{mix}}} + 24n^2 e^{-4r/L_{\mathrm{basis}}} L_{\mathrm{growth}}^{2\log(3x)+4} + 2n^4 e^{-4r/L_{\mathrm{basis}} - x/L_{\mathrm{mix}}})$.

Lemma 9. (Specific Upper Bounds). For $x \geq 1$, $r > 0$, and $g \in \mathcal{G}$,

$$\mathrm{E}[\widehat{\xi}_g^4] \leq f_{\mathrm{SPL}}(r,x), \qquad \mathrm{E}[(G_i'\widehat{\xi})^4] \leq L_{\mathrm{basis}}^4 f_{\mathrm{SPL}}(r,x),$$

$$\mathrm{E}[\tilde{X}_i^4] \leq f_{\mathrm{PAR}}(r,x), \qquad \mathrm{E}[(\tilde{X}_i\varepsilon_i)^4] \leq L_{\mathrm{mom}} f_{\mathrm{PAR}}(r,x),$$

$$\mathrm{E}[(X_i^{\sharp(r)})^4] \leq f_{\mathrm{PAR}}(r,x), \qquad \mathrm{E}[(X_i^{\sharp(r)}\varepsilon_i)^4] \leq L_{\mathrm{mom}} f_{\mathrm{PAR}}(r,x).$$

Proof of Lemma 9. As in the previous lemma, the denominator defining $\widehat{\xi}_g$ has bound $\left(\sum_{i\in S}\tilde{g}(i)^2\right)^{-4} \leq L_{\text{basis}}^8 b^{-4}$. For the numerator, note that for $i \in \mathcal{K}_g(r)$, $|\tilde{g}(i)| \leq L_{\text{basis}}$ by Regularity Condition 7, and $\mathrm{E}[|\tilde{g}(i)X_i|^4] \leq L_{\text{basis}}^4 L_{\text{mom}}$. Apply Lemma 5 to yield that $(|\mathcal{K}_g(r)|^{1/2})^4 \mathrm{E}[(\frac{1}{|\mathcal{K}_g(r)|^{1/2}}\sum_{i\in\mathcal{K}_g(r)} X_i\tilde{g}(i))^4] \leq |\mathcal{K}_g(r)|^2 f_{4\text{TH}}(\mathcal{K}_g(r),x,0) L_{\text{basis}}^4 L_{\text{mom}}$. $\mathrm{E}[(\widehat{\xi}_g^{\sharp(r)})^4] \leq b^{-4} L_{\text{basis}}^{12} L_{\text{mom}} L_{\text{growth}}^{2\log(2r+2z)+4} f_{4\text{TH}}(\mathcal{K}_g(r),x,0)$. Therefore combining the above, as well as the bounds from Lemma 8, gives $\mathrm{E}[\widehat{\xi}_g^4] = \mathrm{E}[(\widehat{\xi}_g^{\flat(r)} + \widehat{\xi}_g^{\sharp(r)})^4] \leq 2^{4-1}(\mathrm{E}[(\widehat{\xi}_g^{\flat(r)})^4] + \mathrm{E}[(\widehat{\xi}_g^{\sharp(r)})^4]) \leq f_{\text{SPL}}(r,x)$. Next, $\mathrm{E}[(G_i'\widehat{\xi})^4] = \mathrm{E}[(\sum_{g:\text{supp}(g)\ni i} g(i)\widehat{\xi}_g)^4] \leq |\{g : \text{supp}(g) \ni i\}|^3 \sum_{g:\text{supp}(g)\ni i} g(i)^4 \mathrm{E}[\widehat{\xi}_g^4]$. Using that $|g(i)| \leq 1$ and that $|\{g : \text{supp}(g) \ni i\}| \leq L_{\text{basis}}$ gives $\mathrm{E}[(G_i'\widehat{\xi})^4] \leq L_{\text{basis}}^4 \max_{g\in\mathcal{G}} \mathrm{E}[\widehat{\xi}_g^4] \leq L_{\text{basis}}^4 f_{\text{SPL}}(r,x)$. Next, $\tilde{X}_i = X_i - G_i'\widehat{\xi}$ so $\mathrm{E}[\tilde{X}_i^4] \leq 2^{4-1}(\mathrm{E}[X_i^4] + \mathrm{E}[(G_i'\widehat{\xi})^4])$. Using $\mathrm{E}[X_i^4] \leq L_{\text{mom}}$ and the previous bound for $\mathrm{E}[(G_i'\widehat{\xi})^4]$ gives the third statement of the lemma. For the fourth statement, note that by the moment conditions, $\mathrm{E}[\varepsilon_i^4\tilde{X}_i^4] \leq \sup_{\mathcal{E}\subseteq\mathcal{B}(\mathbb{R})} \mathrm{E}[\varepsilon_i^4|\tilde{X}_i^4 \in \mathcal{E}]\mathrm{E}[\tilde{X}_i^4] \leq L_{\text{mom}}(8L_{\text{mom}} + 8L_{\text{basis}}^4 f_{\text{SPL}}(r,x)) = L_{\text{mom}} f_{\text{PAR}}(r,x)$. For the fifth and sixth statements, note that $X_i^{\sharp(r)} = X_i - \sum_{g\in\mathcal{G}_i} g(i)\widehat{\xi}_g^{\sharp(r)}$. The same argument used above to bound $\mathrm{E}[\tilde{X}_i^4]$ applies with $\widehat{\xi}_g^{\sharp(r)}$ in place of $\widehat{\xi}_g$ yielding Lemma 9's final two required bounds. ■

Lemma 10. (Specific Lower Bounds). Let $T \subseteq S$ be nonempty. Then

$$\mathrm{E}\Big[\Big(|T|^{-1/2}\sum_{i\in T}\tilde{X}_i\varepsilon_i\Big)^2\Big] \geq f_{\text{COND}}, \mathrm{E}\Big[\Big(|T|^{-1/2}\sum_{i\in T}X_i^{\sharp(r)}\varepsilon_i\Big)^2\Big] \geq f_{\text{COND}},$$
$$\mathrm{E}\Big[|T|^{-1}\sum_{i\in T}\tilde{X}_i^2\Big] \geq f_{\text{COND}}, \mathrm{E}\Big[|T|^{-1}\sum_{i\in T}(X_i^{\sharp(r)})^2\Big] \geq f_{\text{COND}}.$$

Proof of Lemma 10. Let $\widehat{\Omega}_T = |T|^{-1}\sum_{i\in T^2}\varepsilon_{i_1}\tilde{X}_{i_1}\varepsilon_{i_2}\tilde{X}_{i_2}$. Let M_{ij} be such that $\tilde{X}_i = \sum_{j\in S} M_{ij}X_j$. Note, in projection terminology, the matrix with entries M_{ij} is the orthogonal projection onto $\text{span}(\mathcal{G})^\perp$. Note that $\sum_{i\in T}\varepsilon_i\tilde{X}_i = \sum_{i,j\in S} a_{ij}\varepsilon_i X_j$, where $a_{ij} = |T|^{-1/2}1_{i\in T}M_{ij}$. By Regularity Condition 4, $\mathrm{E}[\widehat{\Omega}_T] \geq \frac{1}{L_{\text{cond}}}\sum_{i,j\in S} a_{ij}^2 = \frac{1}{L_{\text{cond}}}\frac{1}{|T|}\sum_{i\in T}\sum_{j\in S} M_{ij}^2$. Because the matrix with entries M_{ij} is an orthogonal projection matrix, $\sum_{j\in S} M_{ij}^2 = M_{ii}$. Therefore $\mathrm{E}[\widehat{\Omega}_T] \geq \frac{1}{L_{\text{cond}}}\frac{1}{|T|}\sum_{i\in T} M_{ii}$. Next, the Frisch-Waugh Theorem gives $\widehat{\xi}_g = D_g^{-1}\sum_{j\in S} X_j\tilde{g}(j)$ where $D_g = \sum_{j\in S}\tilde{g}(j)^2$. Further, $G_i'\widehat{\xi} = \sum_{g\in\mathcal{G}} g(i)\widehat{\xi}_g = \sum_{j\in S}(\sum_{g\in\mathcal{G}} g(i)\tilde{g}(j)D_g^{-1})X_j$. Let $P_{ii} = 1 - M_{ii}$. By Basis Regularity, $D_g \geq b/L_{\text{basis}}^2$. Also by Basis Regularity, $|g(i)| \leq 1, |\tilde{g}(i)| \leq L_{\text{basis}}$, $|\{g : i \in \text{supp}(g)\}| \leq L_{\text{basis}}$, and thus $P_{ii} = \sum_{g\in\mathcal{G}} g(i)\tilde{g}(i)D_g^{-1} \leq \sum_{g:i\in\text{supp}(g)} |g(i)||\tilde{g}(i)|D_g^{-1} \leq \sum_{g:i\in\text{supp}(g)} \frac{L_{\text{basis}}}{b/L_{\text{basis}}^2} \leq \frac{L_{\text{basis}}^2}{b/L_{\text{basis}}^2} \leq \frac{L_{\text{basis}}^4}{b}$. Therefore $\mathrm{E}[(|T|^{-1/2}\sum_{i\in T}\tilde{X}_i\varepsilon_i)^2] \geq \frac{1}{L_{\text{cond}}}(1 - \frac{L_{\text{basis}}^4}{b})$. This is in turn bounded by f_{COND}. Similarly, for fixed i set $c_j = M_{ij}$ so that $\tilde{X}_i = \sum_{j\in S} c_j X_j$. Then similarly to above $\mathrm{E}[\tilde{X}_i^2] \geq \frac{1}{L_{\text{cond}}}(1 - L_{\text{basis}}^4/b)$. Averaging over $i \in T$ gives $\mathrm{E}[|T|^{-1}\sum_{i\in T}\tilde{X}_i^2] \geq f_{\text{COND}}$.

To show the statements containing $X_i^{\sharp(r)}$, define $M_{ij}^{\sharp(r)}$ through $X_i^{\sharp(r)} = \sum_{j\in S} M_{ij}^{\sharp(r)}X_j$. Then $M_{ij}^{\sharp(r)} = 1_{i=j} - \sum_{g\in\mathcal{G}} g(i)1_{j\in\mathcal{K}_g(r)}\tilde{g}(j)D_g^{-1}$. (The difference to M_{ij} is the appearance of $1_{j\in\mathcal{K}_g(r)}$.) Then if $g(i) \neq 0$, then $i \in \mathcal{K}_g(r)$ for all $r > 0$. Therefore $M_{ii}^{\sharp(r)} = 1 - \sum_{g\in\mathcal{G}} g(i)\tilde{g}(i)D_g^{-1} = M_{ii}$. Additionally, though $M_{ij}^{\sharp(r)}$ do not assemble into a idempotent matrix anymore, it still holds that $\sum_{j\in S}(M_{ij}^{\sharp(r)})^2 \geq (M_{ii}^{\sharp(r)})^2$. Therefore, similarly as to above, $\mathrm{E}[(|T|^{-1/2}\sum_{i\in T} X_i^{\sharp(r)}\varepsilon_i)^2] \geq \frac{1}{L_{\text{cond}}}\frac{1}{T}\sum_{i\in T}(M_{ii}^{\sharp(r)})^2 \geq \frac{1}{L_{\text{cond}}}(1 - \frac{L_{\text{basis}}^4}{b})^2 \geq f_{\text{COND}}$. Also $\mathrm{E}[|T|^{-1}\sum_{i\in T}(X_i^{\sharp(r)})^2] \geq f_{\text{COND}}$. ■

Lemma 11. (Specific Laws of Large Number). For $x \geq 1$, $r > 0$, $T \subseteq S$, and $c > 0$,

$$\begin{aligned}
\Pr\Big(\Big|\frac{1}{|T|}\sum_{i\in T}\tilde{X}_i\varepsilon_i\Big| \geq c\Big) &\leq 4c^{-2} f_{\mathrm{LLN}}(T, x, 4z+4r)(L_{\mathrm{mom}} f_{\mathrm{PAR}}(r,x))^{1/2} \\
&\quad + 4c^{-2} L_{\mathrm{basis}}^2 L_{\mathrm{mom}}^{1/2} f_{\mathrm{TRNC}}(r,x)^{1/2}, \\
\Pr\Big(\Big|\frac{1}{|T|}\sum_{i\in T}(\tilde{X}_i^2 - \mathrm{E}[\tilde{X}_i^2])\Big| \geq c\Big) &\leq 9c^{-2} f_{\mathrm{LLN}}(T, x, 4z+4r) f_{\mathrm{PAR}}(r,x) \\
&\quad + 36c^{-2} L_{\mathrm{basis}}^2 (f_{\mathrm{PAR}}(r,x) f_{\mathrm{TRNC}}(r,x))^{1/2} + 9c^{-2} L_{\mathrm{basis}}^4 f_{\mathrm{TRNC}}(r,x).
\end{aligned}$$

Proof of Lemma 11. For the first statement, because $\tilde{X}_i = X_i^{\sharp(r)} + X_i^{\flat(r)}$, the event $|\frac{1}{|T|}\sum_{i\in T}\tilde{X}_i\varepsilon_i| \geq c$ is contained in the union of the events $|\frac{1}{|T|}\sum_{i\in T} X_i^{\sharp(r)}\varepsilon_i| \geq c/2$ and $|\frac{1}{|T|}\sum_{i\in T} X_i^{\flat(r)}\varepsilon_i| \geq c/2$. The former mixes with delay $4z+4r$, hence by Lemma 4, has probability bounded by the first term of the display in the statement of Lemma 11. The later follows from Chebyshev's inequality, using the bound from Lemma 8. The second statement is similar, using the decomposition $\tilde{X}_i^2 = (X_i^{\sharp(r)})^2 + 2X_i^{\sharp(r)}X_i^{\flat(r)} + (X_i^{\flat(r)})^2$ and using three events on which the above corresponding sample averages deviate from their means by more than $c/3$. ■

Let $K_{i_1 i_2} = k(d_{i_1 i_2}/h)$ and $\widehat{\varepsilon}_i = Y_i - X_i\widehat{\beta} - G_i\widehat{\gamma}$. Let $\delta_\beta = \beta_0 - \widehat{\beta}$. Let

$$\begin{aligned}
\widehat{\Omega}^K &= n^{-1}\sum_{i\in S^2} K_{i_1 i_2}\widehat{\varepsilon}_{i_1}\tilde{X}_{i_1}\widehat{\varepsilon}_{i_2}\tilde{X}_{i_2}, \quad \widehat{\Omega}_0^K = n^{-1}\sum_{i\in S^2} K_{i_1 i_2}\varepsilon_{i_1}\tilde{X}_{i_1}\varepsilon_{i_2}\tilde{X}_{i_2}, \\
\widehat{\Omega}_0 &= n^{-1}\sum_{i\in S^2} \varepsilon_{i_1}\tilde{X}_{i_1}\varepsilon_{i_2}\tilde{X}_{i_2}.
\end{aligned}$$

Let

$$\begin{aligned}
f_{\mathrm{VAR}}(r,x) &= n^{-1} L_{\mathrm{growth}}^{3\log 3x+6} 2L_{\mathrm{mom}} f_{\mathrm{PAR}}(r,x) \\
&\quad + 2L_{\mathrm{growth}}^{2\log h+4}(2L_{\mathrm{mom}} f_{\mathrm{PAR}}(r,x) + 2(L_{\mathrm{mom}} f_{\mathrm{PAR}}(r,x))^{1/2})\exp(-(x-(4r+4z))/L_{\mathrm{mix}}) \\
&\quad + 30L_{\mathrm{growth}}^{2\log h+4}(L_{\mathrm{mom}} L_{\mathrm{basis}}^4 f_{\mathrm{TRNC}}(r,x))^{1/4}(L_{\mathrm{mom}} L_{\mathrm{basis}}^4 f_{\mathrm{TRNC}}(r,x) + L_{\mathrm{mom}} f_{\mathrm{PAR}}(r,x))^{3/4}.
\end{aligned}$$

Lemma 12 (Empirical Variance Consistency). For any $x \geq 1$, $r > 0$, and $c > 0$, and for $h \geq 1$, $\Pr(|\widehat{\Omega}_0^K - \mathrm{E}[\widehat{\Omega}_0^K]| \geq c) \leq c^{-2} f_{\mathrm{VAR}}(r,x)$.

Proof of Lemma 12. Define A, C_1, C_2, C_3 as in Lemma 2 and specialize to $y = h$. Use $\nu = 4z + 4r$. Let $W_i = \varepsilon_i X_i^{\sharp(r)}$. For a remainder term $Z^{\flat(r)} = \frac{1}{n^2}\sum_{i\in S^4, V\in\{W,\varepsilon X^{\flat(r)}\}^4 \text{ not all } W} K_{i_1 i_2}K_{i_3 i_4}(V_{i_1}V_{i_2} - \mathrm{E}[V_{i_1}V_{i_2}])(V_{i_3}V_{i_4} - \mathrm{E}[V_{i_3}V_{i_4}])$, the difference $\mathrm{E}[(\widehat{\Omega}_0^K - \mathrm{E}[\widehat{\Omega}_0^K])^2]$ expands to

$$\begin{aligned}
&\mathrm{E}\Big[\frac{1}{n^2}\sum_{i\in S^4} K_{i_1 i_2}K_{i_3 i_4}(W_{i_1}W_{i_2} - \mathrm{E}[W_{i_1}W_{i_2}])(W_{i_3}W_{i_4} - \mathrm{E}[W_{i_3}W_{i_4}])\Big] + \mathrm{E}[Z^{\flat(r)}] \\
&= \frac{1}{n^2}\sum_{i\in A} K_{i_1 i_2}K_{i_3 i_4}\Big(\mathrm{E}[W_{i_1}W_{i_2}W_{i_3}W_{i_4}] - \mathrm{E}[W_{i_1}W_{i_2}]\mathrm{E}[W_{i_3}W_{i_4}]\Big) + \mathrm{E}[Z^{\flat(r)}].
\end{aligned}$$

Let $M_j = \max_{i\in C_j} K_{i_1 i_2}K_{i_3 i_4}|\mathrm{E}[W_{i_1}W_{i_2}W_{i_3}W_{i_4}] - \mathrm{E}[W_{\pi i_1}W_{\pi i_2}]\mathrm{E}[W_{\pi i_3}W_{\pi i_4}]|$, $j \leq 3$, with π being the permutation defining membership into C_2 and C_3, and $\pi = \mathrm{id}$ for the C_1 case. By Lemma 9, $M_1 \leq 2L_{\mathrm{mom}} f_{\mathrm{PAR}}(r,x)$. Next define $M_2 \leq M_{2a} + M_{2b}$ with $M_{2a} = \max_{i\in C_2} K_{i_1 i_2}K_{i_3 i_4}|\mathrm{E}[W_{i_1}W_{i_2}W_{i_3}W_{i_4}] - \mathrm{E}[W_{\pi i_1}]\mathrm{E}[W_{\pi i_2}W_{\pi i_3}W_{\pi i_4}]|$, $M_{2b} = \max_{i\in C_2} K_{i_1 i_2}K_{i_3 i_4}|\mathrm{E}[W_{\pi i_1}]\mathrm{E}[W_{\pi i_2}W_{\pi i_3}W_{\pi i_4}] - \mathrm{E}[W_{i_1}W_{i_2}]\mathrm{E}[W_{i_3}W_{i_4}]|$.

Since $i \in C_2$, there is a permutation π such that $d_{\pi i_1,\{\pi i_2,\pi i_3,\pi i_4\}} \geq x$. Because $\{W_i\}_{i\in S}$ mixes with delay ν by Lemma 8, it follows that $M_{2a} \leq 2L_{\mathrm{mom}} f_{\mathrm{PAR}}(r,x)\exp(-(x-\nu)/L_{\mathrm{mix}})$. In addition, $\mathrm{E}[W_{\pi i_1}] = 0$ and either the bound $|\mathrm{E}[W_{i_1}W_{i_2}]| = |\mathrm{E}[W_{i_1}W_{i_2}] - \mathrm{E}[W_{i_1}]\mathrm{E}[W_{i_2}]| \leq 2(L_{\mathrm{mom}} f_{\mathrm{PAR}}(r,x))^{1/2}\exp(-(x-\nu)/L_{\mathrm{mix}})$ holds or the same bound for (i_3,i_4) holds. Together, these imply $M_{2b} \leq 2(L_{\mathrm{mom}} f_{\mathrm{PAR}}(r,x))^{1/2}\exp(-(x-\nu)/L_{\mathrm{mix}})$. For M_3, $|\mathrm{E}[W_{i_1}W_{i_2}W_{i_3}W_{i_4}] - \mathrm{E}[W_{\pi i_1}W_{\pi i_2}]\mathrm{E}[W_{\pi i_3}W_{\pi i_4}]| \leq 2L_{\mathrm{mom}} f_{\mathrm{PAR}}(r,x)\exp(-(x-\nu)/L_{\mathrm{mix}})$. Either $\mathrm{E}[W_{\pi i_1}W_{\pi i_2}]\mathrm{E}[W_{\pi i_3}W_{\pi i_4}] = \mathrm{E}[W_{i_1}W_{i_2}]\mathrm{E}[W_{i_3}W_{i_4}]$ and $|\mathrm{E}[W_{i_1}W_{i_2}W_{i_3}W_{i_4}] - \mathrm{E}[W_{i_1}W_{i_2}]\mathrm{E}[W_{i_3}W_{i_4}]| \leq 2L_{\mathrm{mom}} f_{\mathrm{PAR}}(r,x)\exp(-(x-\nu)/L_{\mathrm{mix}})$ or $d_{\pi i_1 \pi i_2} \geq x$ and $d_{\pi i_3 \pi i_4} \geq x$ giving that both $|\mathrm{E}[W_{i_1}W_{i_2}]| \leq 2L_{\mathrm{mom}}^{1/2} f_{\mathrm{PAR}}(r,x)^{1/2}\exp(-(x-\nu)/L_{\mathrm{mix}})$ and $|\mathrm{E}[W_{i_3}W_{i_4}]| \leq 2L_{\mathrm{mom}}^{1/2} f_{\mathrm{PAR}}(r,x)^{1/2}\exp(-(x-\nu)/L_{\mathrm{mix}})$. The bound for M_3 is then $\exp(-(x-\nu)/L_{\mathrm{mix}})(2L_{\mathrm{mom}}^{1/2} f_{\mathrm{PAR}}(r,x)^{1/2} + 2L_{\mathrm{mom}} f_{\mathrm{PAR}}(r,x))$.

Additionally, for $Z^{\flat(r)}$, there are $2^4 - 1 = 15$ ways to take a combination of 4 $\sharp(r)$ and $\flat(r)$-s, without all $\sharp(r)$-s. For each such combination $|\mathrm{E}[(V_{i_1}V_{i_2} - \mathrm{E}[V_{i_1}V_{i_2}])(V_{i_3}V_{i_4} - \mathrm{E}[V_{i_3}V_{i_4}])]|$ is bounded by the sum of the two terms $\mathrm{E}[|V_{i_1}V_{i_2}V_{i_3}V_{i_4}|]$ and $\mathrm{E}[|V_{i_1}V_{i_2}|]\mathrm{E}[|V_{i_3}V_{i_4}|]$. This gives 30 such terms, all with a common Hölder bound $\max_{i\in S}\mathrm{E}[(X_i^{\flat(r)}\varepsilon_i)^4]^{1/4}(\max_{i\in S}\max(\mathrm{E}[(X_i^{\flat(r)}\varepsilon_i)^4], \mathrm{E}[(X_i^{\sharp(r)}\varepsilon_i)^4]))^{3/4}$, and thus $|\mathrm{E}[Z^{\flat(r)}]| \leq 30L_{\mathrm{growth}}^{2\log h+4}(L_{\mathrm{mom}}L_{\mathrm{basis}}^4 f_{\mathrm{TRNC}}(r,x))^{1/4}(L_{\mathrm{mom}}L_{\mathrm{basis}}^4 f_{\mathrm{TRNC}}(r,x) + L_{\mathrm{mom}} f_{\mathrm{PAR}}(r,x))^{3/4}$.

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

$$\begin{aligned}\mathrm{E}[(\widehat{\Omega}_0^K - \mathrm{E}[\widehat{\Omega}_0^K])^2] &\leq \frac{1}{n^2}|C_1|M_1 + \frac{1}{n^2}|A|M_2 + \frac{1}{n^2}|A|M_3 + |\mathrm{E}[Z^{\flat(r)}]| \\ &\leq f_{\mathrm{VAR}}(r,x).\end{aligned}$$

Using Chebyshev's inequality gives the lemma. ■

Next let $\delta_\beta = \beta_0 - \widehat{\beta}$. Denote $\mathrm{E}_S = n^{-1}\sum_{i_1\in S}$ and $\mathrm{E}_K = \sum_{i_2\in B_h(i_1)} K_{i_1 i_2}$ and additionally $\mathrm{E}_S\mathrm{E}_K = n^{-1}\sum_{i_1\in S}\mathrm{E}_K\sum_{i_2\in B_h(i_1)} K_{i_1 i_2}$. Then noting that $\widehat{\gamma} = \widehat{\xi}\delta_\beta + \widehat{\eta}$ gives $\widehat{\varepsilon}_i = \varepsilon_i + X_i\delta_\beta - G_i'\widehat{\gamma} = \varepsilon_i + \tilde{X}_i\delta_\beta - G_i'\widehat{\eta}$ and thus

$$\begin{aligned}\widehat{\Omega}^K - \widehat{\Omega}_0^K = {}& \mathrm{E}_S\mathrm{E}_K\tilde{X}_{i_1}(\varepsilon_{i_1} + \tilde{X}_{i_1}\delta_\beta - G_{i_1}'\widehat{\eta})\tilde{X}_{i_2}(\varepsilon_{i_2} + \tilde{X}_{i_2}\delta_\beta - G_{i_2}'\widehat{\eta}) \\ & - \mathrm{E}_S\mathrm{E}_K\tilde{X}_{i_1}\varepsilon_{i_1}\tilde{X}_{i_2}\varepsilon_{i_2}.\end{aligned}$$

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

$$\begin{aligned}\delta_1(u) &= \mathrm{E}_S\mathrm{E}_K\ \tilde{X}_{i_1}\tilde{X}_{i_1}u\tilde{X}_{i_2}\tilde{X}_{i_2}u, \quad \delta_2(u) = -2\mathrm{E}_S\mathrm{E}_K\ \tilde{X}_{i_1}G_{i_1}'\widehat{\eta}\tilde{X}_{i_2}\tilde{X}_{i_2}u, \\ \delta_3(u) &= \mathrm{E}_S\mathrm{E}_K\ \tilde{X}_{i_1}G_{i_1}'\widehat{\eta}\tilde{X}_{i_2}G_{i_2}'\widehat{\eta}, \quad \delta_4(u) = 2\mathrm{E}_S\mathrm{E}_K\ \tilde{X}_{i_1}\varepsilon_{i_1}\tilde{X}_{i_2}\tilde{X}_{i_2}u, \\ \delta_5(u) &= 2\mathrm{E}_S\mathrm{E}_K\ \tilde{X}_{i_1}\varepsilon_{i_1}\tilde{X}_{i_2}G_{i_2}'\widehat{\eta}.\end{aligned}$$

In the special (also random) case $u = \delta_\beta$, the decomposition $\widehat{\Omega}^K - \widehat{\Omega}_0^K = \delta_1(\delta_\beta) + ... + \delta_5(\delta_\beta)$ holds.

Let

$$
\begin{aligned}
f_{\mathrm{BETA}}(r,x,u) &= 4f_{\mathrm{LLN}}(S,x,4z+4r)(L_{\mathrm{mom}}f_{\mathrm{PAR}}(r,x))^{1/2} + 4L_{\mathrm{basis}}^2 L_{\mathrm{mom}}^{1/2} f_{\mathrm{TRNC}}(r,x)^{1/2} \\
&\quad + u^2(9f_{\mathrm{LLN}}(S,x,4z+4r)f_{\mathrm{PAR}}(r,x) + 36L_{\mathrm{basis}}^2(f_{\mathrm{PAR}}(r,x)f_{\mathrm{TRNC}}(r,x))^{1/2} + 9L_{\mathrm{basis}}^4 f_{\mathrm{TRNC}}(r,x)), \\
f_{\mathrm{RES},0}(r,x) &= 16L_{\mathrm{basis}}^2 f_{\mathrm{PAR}}(r,x)\big(f_{\mathrm{SPL}}(r,x)f_{\mathrm{TRNC}}(r,x)\big)^{1/2} + 4L_{\mathrm{basis}}^2 f_{\mathrm{PAR}}(r,x)f_{\mathrm{TRNC}}(r,x) \\
&\quad + 16L_{\mathrm{basis}}^4 f_{\mathrm{SPL}}(r,x)\big(f_{\mathrm{PAR}}(r,x)f_{\mathrm{TRNC}}(r,x)\big)^{1/2} + 4L_{\mathrm{basis}}^8 f_{\mathrm{SPL}}(r,x)f_{\mathrm{TRNC}}(r,x), \\
f_{\mathrm{RES},1}(r,x) &= 4b^{-4}L_{\mathrm{kern}}^2 L_{\mathrm{basis}}^{16} L_{\mathrm{mom}}\, x^2 h^{-2} L_{\mathrm{growth}}^{4\log x+8} f_{\mathrm{PAR}}(r,x) \\
&\quad \times \Big(n^{-1}L_{\mathrm{growth}}^{3\log\left(3(3x+4r+4z)\right)+6} + 2L_{\mathrm{growth}}^{2\log(h+2x)+4}\exp\big(-x/L_{\mathrm{mix}}\big)\Big) \\
&\quad + 16b^{-4}L_{\mathrm{kern}}^2 L_{\mathrm{basis}}^{16} h^{-2}\Big(L_{\mathrm{mom}}f_{\mathrm{PAR}}(r,x)\exp\big(-2x/L_{\mathrm{basis}}\big) + L_{\mathrm{growth}}^{2\log x+4} f_{\mathrm{RES},0}(r,x)\Big), \\
f_{\mathrm{RES},2}(r,x) &= 4b^{-2}L_{\mathrm{kern}}^2 L_{\mathrm{basis}}^{8} L_{\mathrm{mom}} x^2 h^{-2} L_{\mathrm{growth}}^{2\log x+4} f_{\mathrm{PAR}}(r,x) \\
&\quad \times \Big(n^{-1}L_{\mathrm{growth}}^{3\log(3(3x+4r+4z))+6} + 2L_{\mathrm{growth}}^{2\log(h+x)+4}\exp(-x/L_{\mathrm{mix}})\Big) \\
&\quad + 16b^{-2}L_{\mathrm{kern}}^2 L_{\mathrm{basis}}^{8} h^{-2}\Big(L_{\mathrm{mom}}f_{\mathrm{PAR}}(r,x)\exp(-2x/L_{\mathrm{basis}}) + L_{\mathrm{growth}}^{2\log x+4} f_{\mathrm{RES},0}(r,x)\Big).
\end{aligned}
$$

Finally,

$$
\begin{aligned}
f_{\mathrm{RES}}(r,x,u) &= L_{\mathrm{growth}}^{\log h+2}\Big(u^2 f_{\mathrm{PAR}}(r,x) + 2|u|f_{\mathrm{PAR}}(r,x)^{3/4}\big(L_{\mathrm{basis}}f_{\mathrm{SPL}}(r,x)^{1/4} + L_{\mathrm{mom}}^{1/4}\big)\Big) \\
&\quad + 4f_{\mathrm{RES},1}(r,x)^{1/2} + 4f_{\mathrm{RES},2}(r,x)^{1/2}.
\end{aligned}
$$

Lemma 13 (Parameter Estimate Consistency). For $u>0$, $x\geq 1$, $r>0$, if $b > L_{\mathrm{basis}}^4$,

$$\Pr(\delta_\beta^2 \geq u^2) \leq 4u^{-2} f_{\mathrm{COND}}^{-2} f_{\mathrm{BETA}}(r,x,u)$$

Proof of Lemma 13. Let $\widehat{Q} = n^{-1}\sum_{i\in S}\tilde{X}_i^2$ and $Q_0 = \mathrm{E}[\widehat{Q}]$. $\Pr(\delta_\beta^2 \geq u^2) \leq \Pr(|n^{-1}\sum_{i\in S}\tilde{X}_i\varepsilon_i| \geq uQ_0/2) + \Pr(\widehat{Q} \leq Q_0/2)$ Simplifying and applying Lemma 11 together with the f_{COND} lower bound of Lemma 10 for Q_0 gives Lemma 13. ■

Lemma 14 (Residual Estimate Replacement Bounds). For $j=1,...,5$, $x\geq 1, y\geq 1, r>0, 0<c<1$ and $u\in\mathbb{R}$, $\Pr(|\delta_j(u)|\geq c) \leq c^{-1} f_{\mathrm{RES}}(r,x,u)$.

Proof of Lemma 14. For $\delta_1(u)$, Let U_{i_1} be the term inside the E_S operation, so that $U_{i_1} = \mathrm{E}_K \tilde{X}_{i_1}^2 \tilde{X}_{i_2}^2 u^2$. Then $|K_{i_1 i_2}| \leq 1$ and by Hölder's inequality, $\mathrm{E}[|U_{i_1}|] \leq L_{\mathrm{growth}}^{\log h+2} f_{\mathrm{PAR}}(r,x)u^2$. Then the first moment bound is $\mathrm{E}[|\delta_1(u)|] \leq L_{\mathrm{growth}}^{\log h+2} f_{\mathrm{PAR}}(r,x)u^2$ and then Markov's inequality gives $\Pr(|\delta_1(u)|\geq c) \leq c^{-1}\mathrm{E}[|\delta_1(u)|] \leq c^{-1}u^2 L_{\mathrm{growth}}^{\log h+2} f_{\mathrm{PAR}}(r,x)$. The remaining terms are then bounded analogously. Note that $\mathrm{E}[(G_i'\widehat{\eta})^4] \leq L_{\mathrm{basis}}^4 f_{\mathrm{SPL}}(r,x)$ by an identical bounding argument as made for $\mathrm{E}[(G_i'\widehat{\xi})^4]$. Summarizing thus far, the following bounds hold.

$$
\begin{aligned}
&\Pr(|\delta_1(u)|\geq c) \leq c^{-1}u^2 L_{\mathrm{growth}}^{\log h+2} f_{\mathrm{PAR}}(r,x) \\
&\Pr(|\delta_2(u)|\geq c) \leq c^{-1}|u| 2L_{\mathrm{growth}}^{\log h+2} L_{\mathrm{basis}} f_{\mathrm{SPL}}(r,x)^{1/4} f_{\mathrm{PAR}}(r,x)^{3/4}, \\
&\Pr(|\delta_4(u)|\geq c) \leq c^{-1}|u| 2L_{\mathrm{growth}}^{\log h+2} L_{\mathrm{mom}}^{1/4} f_{\mathrm{PAR}}(r,x)^{3/4}.
\end{aligned}
$$

For δ_3, write $\widehat{\eta}_g = D_g^{-1}\sum_{j\in S}\tilde{g}(j)\varepsilon_j$, where $D_g = \sum_{m\in S}\tilde{g}(m)^2$, and $H_{ij} = \sum_{g\in\mathcal{G}} g(i)\tilde{g}(j)D_g^{-1}$ and $G_i'\widehat{\eta} = \sum_{j\in S} H_{ij}\varepsilon_j$. Let $\Psi_{j_1 j_2} = \sum_{i_1,i_2\in S} K_{i_1 i_2}\tilde{X}_{i_1}\tilde{X}_{i_2} H_{i_1 j_1} H_{i_2 j_2}$. Therefore

$$\delta_3 = n^{-1}\sum_{j_1,j_2\in S}\varepsilon_{j_1}\varepsilon_{j_2}\Psi_{j_1 j_2}.$$

Now $\sum_{i\in S}\tilde{X}_i H_{ij} = \sum_{g\in\mathcal{G}}\tilde{g}(j)D_g^{-1}\sum_{i\in S}\tilde{X}_i g(i) = 0$, by the least squares normal equations. Hence, after subtracting the constant $K_{j_1j_2}$ inside the double sum,

$$\Psi_{j_1j_2} = \sum_{i_1,i_2\in S}\big(K_{i_1i_2} - K_{j_1j_2}\big)\tilde{X}_{i_1}\tilde{X}_{i_2}H_{i_1j_1}H_{i_2j_2}.$$

Next define $H_{ij}^{\sharp(x)} = H_{ij}1_{d_{ij}\le x}$, $\Psi_{j_1j_2}^{\sharp(x)} = \sum_{i_1,i_2\in S}\big(K_{i_1i_2} - K_{j_1j_2}\big)X_{i_1}^{\sharp(r)}X_{i_2}^{\sharp(r)}H_{i_1j_1}^{\sharp(x)}H_{i_2j_2}^{\sharp(x)}$, and $\delta_3^{\sharp(x)} = n^{-1}\sum_{j_1,j_2\in S}\varepsilon_{j_1}\varepsilon_{j_2}\Psi_{j_1j_2}^{\sharp(x)}$. Then define $\delta_3^{\flat(x)} = \delta_3 - \delta_3^{\sharp(x)}$. First bound $\delta_3^{\sharp(x)}$. By Basis Regularity, $D_g \ge bL_{\text{basis}}^{-2}$, so if $g(i)\neq 0$, then $i\in\text{supp}(g)$ and as a result $|g(i)\tilde{g}(j)D_g^{-1}| \le b^{-1}L_{\text{basis}}^3\exp(-d_{j\,\text{supp}(g)}/L_{\text{basis}}) \le b^{-1}L_{\text{basis}}^3\exp(-d_{ij}/L_{\text{basis}})$. Since at most L_{basis} functions g satisfy $g(i)\neq 0$, then there is $|H_{ij}| \le b^{-1}L_{\text{basis}}^4\exp(-d_{ij}/L_{\text{basis}})$. Also, if $H_{i_1j_1}^{\sharp(x)}H_{i_2j_2}^{\sharp(x)}\neq 0$, then $d_{i_1j_1}\le x$ and $d_{i_2j_2}\le x$. By Kernel Regularity, $|K_{i_1i_2} - K_{j_1j_2}| \le L_{\text{kern}}h^{-1}(d_{i_1j_1} + d_{i_2j_2}) \le 2L_{\text{kern}}h^{-1}x$. Hence $|\Psi_{j_1j_2}^{\sharp(x)}| \le 2b^{-2}L_{\text{kern}}L_{\text{basis}}^8 x h^{-1}\Upsilon_{j_1}\Upsilon_{j_2}$, where $\Upsilon_j = \sum_{i\in B_x(j)}|X_i^{\sharp(r)}|$. By Lemma 1, $|B_x(j)| \le L_{\text{growth}}^{\log x+2}$, so $\mathrm{E}[\Upsilon_j^4] \le |B_x(j)|^4\max_{i\in S}\mathrm{E}[(X_i^{\sharp(r)})^4] \le L_{\text{growth}}^{4\log x+8}f_{\text{PAR}}(r,x)$. And so

$$\max_{j_1,j_2\in S}\mathrm{E}\big[(\varepsilon_{j_1}\varepsilon_{j_2}\Psi_{j_1j_2}^{\sharp(x)})^2\big] \le 4b^{-4}L_{\text{kern}}^2L_{\text{basis}}^{16}L_{\text{mom}}x^2h^{-2}L_{\text{growth}}^{4\log x+8}f_{\text{PAR}}(r,x).$$

Now $\Psi_{j_1j_2}^{\sharp(x)} = 0$ whenever $d_{j_1j_2} > h + 2x$. Indeed, if a summand in the definition of $\Psi_{j_1j_2}^{\sharp(x)}$ is nonzero, then $d_{i_1j_1}\le x$, $d_{i_2j_2}\le x$, and $K_{i_1i_2}\neq 0$, which implies $d_{i_1i_2}\le h$. By the triangle inequality, $d_{j_1j_2} \le d_{j_1i_1} + d_{i_1i_2} + d_{i_2j_2} \le h + 2x$. Therefore only indices in the set $A = \{j\in S^4 : d_{j_1j_2}\le h+2x,\ d_{j_3j_4}\le h+2x\}$ contribute to $\mathrm{E}[(\delta_3^{\sharp(x)})^2]$. Decompose this set as in Lemma 2 using separation parameter $3x+4r+4z$. For $j\in C_2\cup C_3$, the two factors $\varepsilon_{j_1}\varepsilon_{j_2}\Psi_{j_1j_2}^{\sharp(x)}$ and $\varepsilon_{j_3}\varepsilon_{j_4}\Psi_{j_3j_4}^{\sharp(x)}$ depend only on observations in the sets $B_{x+2r+2z}(j_1)\cup B_{x+2r+2z}(j_2)$ and $B_{x+2r+2z}(j_3)\cup B_{x+2r+2z}(j_4)$, respectively. This is because $H_{ij}^{\sharp(x)}\neq 0$ forces $d_{ij}\le x$, and by Lemma 8, $X_i^{\sharp(r)}$ depends only on $\{X_{i'}\}_{i'\in B_{2r+2z}(i)}$. Hence these two factors mix with delay $2x+4r+4z$. Since the separation parameter in Lemma 2 has been chosen to be $3x+4r+4z$, the mixing bound contributes the factor $\exp(-x/L_{\text{mix}})$ on $C_2\cup C_3$. On the other hand, on C_1 we simply use the second-moment bound displayed above. Therefore, exactly as in Lemma 12,

$$\begin{aligned}\mathrm{E}[(\delta_3^{\sharp(x)})^2] &\le 4b^{-4}L_{\text{kern}}^2L_{\text{basis}}^{16}L_{\text{mom}}x^2h^{-2}L_{\text{growth}}^{4\log x+8}f_{\text{PAR}}(r,x)\\ &\quad\times\Big(n^{-1}L_{\text{growth}}^{3\log(3(3x+4r+4z))+6} + 2L_{\text{growth}}^{2\log(h+2x)+4}\exp(-x/L_{\text{mix}})\Big).\end{aligned}$$

It remains to bound $\delta_3^{\flat(x)} = \delta_3 - \delta_3^{\sharp(x)}$. By construction, every term in $\delta_3^{\flat(x)}$ contains either at least one factor $H_{ij} - H_{ij}^{\sharp(x)}$ or at least one factor $\tilde{X}_i - X_i^{\sharp(r)} = X_i^{\flat(r)}$. For the first type of term, the bound on H_{ij} above gives $|H_{ij} - H_{ij}^{\sharp(x)}| \le b^{-1}L_{\text{basis}}^4\exp(-x/L_{\text{basis}})$. Using this bound, together with the kernel Lipschitz bound and the same moment estimates as above, the contribution to $\mathrm{E}[(\delta_3^{\flat(x)})^2]$ from all terms containing at least one factor $H_{ij} - H_{ij}^{\sharp(x)}$ is bounded by $16b^{-4}L_{\text{kern}}^2L_{\text{basis}}^{16}h^{-2}L_{\text{mom}}f_{\text{PAR}}(r,x)\exp(-2x/L_{\text{basis}})$. For the second type of term, at least one factor $X_i^{\flat(r)}$ appears. Bounding these terms exactly as in the mixed sharp-flat expansions used earlier, and using Lemmas 8 and 9 together with the cardinality bound from Lemma 1 for the x-neighborhoods generated by the sharp H factors, gives the contribution $16b^{-4}L_{\text{kern}}^2L_{\text{basis}}^{16}h^{-2}L_{\text{growth}}^{2\log x+4}f_{\text{RES},0}(r,x)$. Combining the two kinds of terms yields

$$\mathrm{E}[(\delta_3^{\flat(x)})^2] \le 16b^{-4}L_{\text{kern}}^2L_{\text{basis}}^{16}h^{-2}\Big(L_{\text{mom}}f_{\text{PAR}}(r,x)\exp(-2x/L_{\text{basis}}) + L_{\text{growth}}^{2\log x+4}f_{\text{RES},0}(r,x)\Big).$$

As $\delta_3 = \delta_3^{\sharp(x)} + \delta_3^{\flat(x)}$, the union bound, Markov's inequality, Cauchy-Schwarz, and $a^{1/2}+b^{1/2}\le 2(a+b)^{1/2}$ give $\Pr(|\delta_3|\ge c) \le \Pr(|\delta_3^{\sharp(x)}|\ge c/2) + \Pr(|\delta_3^{\flat(x)}|\ge c/2) \le \frac{2}{c}\mathrm{E}[|\delta_3^{\sharp(x)}|] + \frac{2}{c}\mathrm{E}[|\delta_3^{\flat(x)}|] \le \frac{2}{c}\mathrm{E}[(\delta_3^{\sharp(x)})^2]^{1/2} + \frac{2}{c}\mathrm{E}[(\delta_3^{\flat(x)})^2]^{1/2}$ and therefore

$$\Pr(|\delta_3|\ge c) \le 4c^{-1}f_{\text{RES},1}(r,x)^{1/2}.$$

Next bound δ_5. Note $\delta_5 = 2n^{-1}\sum_{i_1,i_2,j\in S} K_{i_1i_2}\tilde{X}_{i_1}\varepsilon_{i_1}\tilde{X}_{i_2}H_{i_2j}\varepsilon_j$. Redefine $\Psi_{i_1j} = \tilde{X}_{i_1}\sum_{i_2\in S} K_{i_1i_2}\tilde{X}_{i_2}H_{i_2j}$, so that $\delta_5 = 2n^{-1}\sum_{i_1,j\in S}\varepsilon_{i_1}\varepsilon_j\Psi_{i_1j}$. Again, $\sum_{i_2\in S}\tilde{X}_{i_2}H_{i_2j} = \sum_{g\in\mathcal{G}}\tilde{g}(j)D_g^{-1}\sum_{i_2\in S}\tilde{X}_{i_2}g(i_2) = 0$ by the normal equations. Subtracting K_{i_1j} inside the sum over i_2 gives $\Psi_{i_1j} = \tilde{X}_{i_1}\sum_{i_2\in S}(K_{i_1i_2}-K_{i_1j})\tilde{X}_{i_2}H_{i_2j}$.

Define $\Psi^{\sharp(x)}_{i_1j} = X^{\sharp(r)}_{i_1}\sum_{i_2\in S}(K_{i_1i_2}-K_{i_1j})X^{\sharp(r)}_{i_2}H^{\sharp(x)}_{i_2j}$, $\delta_5^{\sharp(x)} = 2n^{-1}\sum_{i_1,j\in S}\varepsilon_{i_1}\varepsilon_j\Psi^{\sharp(x)}_{i_1j}$, $\delta_5^{\flat(x)} = \delta_5 - \delta_5^{\sharp(x)}$.

First bound $\delta_5^{\sharp(x)}$. As shown above in the proof of the bound for δ_3, $|H_{ij}| \le b^{-1}L_{\mathrm{basis}}^4\exp(-d_{ij}/L_{\mathrm{basis}})$. Also, if $H^{\sharp(x)}_{i_2j}\neq 0$, then $d_{i_2j}\le x$, and so by Kernel Regularity, $|K_{i_1i_2}-K_{i_1j}| \le L_{\mathrm{kern}}h^{-1}d_{i_2j} \le L_{\mathrm{kern}}xh^{-1}$. Therefore $|\Psi^{\sharp(x)}_{i_1j}| \le b^{-1}L_{\mathrm{kern}}L_{\mathrm{basis}}^4 xh^{-1}|X^{\sharp(r)}_{i_1}|\sum_{i\in B_x(j)}|X^{\sharp(r)}_i|$. By Lemma 1, $|B_x(j)|\le L_{\mathrm{growth}}^{\log x+2}$, and by Lemma 9, $\max_{i\in S}\mathrm{E}[(X^{\sharp(r)}_i)^4]\le f_{\mathrm{PAR}}(r,x)$. Hence $\mathrm{E}[(\sum_{i\in B_x(j)}|X^{\sharp(r)}_i|)^4]\le |B_x(j)|^4\max_{i\in S}\mathrm{E}[(X^{\sharp(r)}_i)^4]\le L_{\mathrm{growth}}^{4\log x+8}f_{\mathrm{PAR}}(r,x)$. It follows by Hölder's inequality and Lemma 9 that

$$\begin{aligned}\max_{i_1,j\in S}\mathrm{E}\big[(\varepsilon_{i_1}\varepsilon_j\Psi^{\sharp(x)}_{i_1j})^2\big] &\le b^{-2}L_{\mathrm{kern}}^2L_{\mathrm{basis}}^8x^2h^{-2}\mathrm{E}\big[\varepsilon_{i_1}^2\varepsilon_j^2(X^{\sharp(r)}_{i_1})^2\Gamma_j^2\big]\\ &\le b^{-2}L_{\mathrm{kern}}^2L_{\mathrm{basis}}^8x^2h^{-2}\mathrm{E}[(X^{\sharp(r)}_{i_1}\varepsilon_{i_1})^4]^{1/2}\mathrm{E}[\varepsilon_j^4]^{1/2}\mathrm{E}[\Gamma_j^4]^{1/2}\\ &\le b^{-2}L_{\mathrm{kern}}^2L_{\mathrm{basis}}^8L_{\mathrm{mom}}x^2h^{-2}L_{\mathrm{growth}}^{2\log x+4}f_{\mathrm{PAR}}(r,x).\end{aligned}$$

Now $\Psi^{\sharp(x)}_{i_1j}=0$ whenever $d_{i_1j}>h+x$. Indeed, if a summand in the definition of $\Psi^{\sharp(x)}_{i_1j}$ is nonzero, then $d_{i_2j}\le x$ and either $K_{i_1i_2}\neq 0$ or $K_{i_1j}\neq 0$. In the first case $d_{i_1i_2}\le h$, and hence by triangle inequality $d_{i_1j}\le d_{i_1i_2}+d_{i_2j}\le h+x$. In the second case $d_{i_1j}\le h$. Thus only indices in the set

$$A = \{i\in S^4 : d_{i_1i_2}\le h+x,\ d_{i_3i_4}\le h+x\}$$

contribute to $\mathrm{E}[(\delta_5^{\sharp(x)})^2]$, where the pair (i_1,i_2) corresponds to the index pair (i_1,j) above and (i_3,i_4) corresponds to the second copy.

Decompose this set as in Lemma 2 using separation parameter $3x+4r+4z$. For $i\in C_2\cup C_3$, the two factors $\varepsilon_{i_1}\varepsilon_{i_2}\Psi^{\sharp(x)}_{i_1i_2}$ and $\varepsilon_{i_3}\varepsilon_{i_4}\Psi^{\sharp(x)}_{i_3i_4}$ depend only on observations in $B_{2r+2z}(i_1)\cup B_{x+2r+2z}(i_2)$ and $B_{2r+2z}(i_3)\cup B_{x+2r+2z}(i_4)$, respectively. This is because $H^{\sharp(x)}_{i_2j}\neq 0$ forces $d_{i_2j}\le x$, and by Lemma 8, each $X^{\sharp(r)}_i$ depends only on $\{X_{i'}\}_{i'\in B_{2r+2z}(i)}$. Hence these two factors mix with delay $2x+4r+4z$. Since the separation parameter in Lemma 2 is $3x+4r+4z$, the mixing bound contributes the factor $\exp(-x/L_{\mathrm{mix}})$ on $C_2\cup C_3$. On the other hand, on C_1 we use the second-moment bound displayed above. Therefore, exactly as in Lemma 12,

$$\begin{aligned}\mathrm{E}[(\delta_5^{\sharp(x)})^2] \le{}& 4b^{-2}L_{\mathrm{kern}}^2L_{\mathrm{basis}}^8L_{\mathrm{mom}}x^2h^{-2}L_{\mathrm{growth}}^{2\log x+4}f_{\mathrm{PAR}}(r,x)\\ &\times\Big(n^{-1}L_{\mathrm{growth}}^{3\log(3(3x+4r+4z))+6}+2L_{\mathrm{growth}}^{2\log(h+x)+4}\exp(-x/L_{\mathrm{mix}})\Big).\end{aligned}$$

It remains to bound $\delta_5^{\flat(x)} = \delta_5-\delta_5^{\sharp(x)}$. By construction, every term in $\delta_5^{\flat(x)}$ contains either at least one factor $H_{ij}-H^{\sharp(x)}_{ij}$ or at least one factor $\tilde{X}_i - X^{\sharp(r)}_i = X^{\flat(r)}_i$. For the first type of term, the bound on H_{ij} above gives $|H_{ij}-H^{\sharp(x)}_{ij}|\le b^{-1}L_{\mathrm{basis}}^4\exp(-x/L_{\mathrm{basis}})$. Using this bound, together with the kernel Lipschitz bound and the same moment estimates as above, the contribution to $\mathrm{E}[(\delta_5^{\flat(x)})^2]$ from all terms containing at least one factor $H_{ij}-H^{\sharp(x)}_{ij}$ is bounded by $16b^{-2}L_{\mathrm{kern}}^2L_{\mathrm{basis}}^8h^{-2}L_{\mathrm{mom}}f_{\mathrm{PAR}}(r,x)\exp(-2x/L_{\mathrm{basis}})$. For the second type of term, at least one factor $X^{\flat(r)}_i$ appears. Bounding these terms exactly as in

the mixed sharp-flat expansions used earlier, and using Lemmas 8 and 9 together with the cardinality bound from Lemma 1 for the x-neighborhoods generated by the sharp H factors, gives the contribution $16b^{-2}L_{\mathrm{kern}}^{2}L_{\mathrm{basis}}^{8}h^{-2}L_{\mathrm{growth}}^{2\log x+4}f_{\mathrm{RES},0}(r,x)$. Combining the two kinds of terms yields

$$\mathrm{E}[(\delta_5^{\flat(x)})^2] \le 16b^{-2}L_{\mathrm{kern}}^{2}L_{\mathrm{basis}}^{8}h^{-2}\Big(L_{\mathrm{mom}}f_{\mathrm{PAR}}(r,x)\exp(-2x/L_{\mathrm{basis}}) + L_{\mathrm{growth}}^{2\log x+4}f_{\mathrm{RES},0}(r,x)\Big).$$

As $\delta_5 = \delta_5^{\sharp(x)} + \delta_5^{\flat(x)}$, union bound, Markov's inequality, Cauchy-Schwarz, and $a^{1/2}+b^{1/2} \le 2(a+b)^{1/2}$ give $\Pr(|\delta_5| \ge c) \le \Pr(|\delta_5^{\sharp(x)}| \ge c/2) + \Pr(|\delta_5^{\flat(x)}| \ge c/2) \le \frac{2}{c}\mathrm{E}[(\delta_5^{\sharp(x)})^2]^{1/2} + \frac{2}{c}\mathrm{E}[(\delta_5^{\flat(x)})^2]^{1/2}$, and therefore

$$\Pr(|\delta_5| \ge c) \le 4c^{-1}f_{\mathrm{RES},2}(r,x)^{1/2}.$$

This concludes the proof of the lemma. ■

Let

$$\begin{aligned} f_{\mathrm{KER}}(r,x) = {} & (L_{\mathrm{kern}}L_{\mathrm{growth}}^{\log x+2}(x/h) + 2n\exp(-(x-(4r+4z))/L_{\mathrm{mix}}))L_{\mathrm{mom}}^{1/2}f_{\mathrm{PAR}}(r,x)^{1/2}. \\ & + 2nL_{\mathrm{mom}}^{1/2}L_{\mathrm{basis}}f_{\mathrm{TRNC}}(r,x)^{1/4}f_{\mathrm{PAR}}(r,x)^{1/4} + nL_{\mathrm{mom}}^{1/2}L_{\mathrm{basis}}^{2}f_{\mathrm{TRNC}}(r,x)^{1/2}. \end{aligned}$$

Lemma 15 (Kernel Approximation). Let $x \ge 1$ and $r > 0$. Then

$$\Big|\frac{1}{n}\mathrm{var}(\sum_{i\in S}\varepsilon_i X_i^{\sharp(r)}) - \mathrm{E}[\widehat{\Omega}_0^K]\Big| \le f_{\mathrm{KER}}(r,x).$$

Proof of Lemma 15. Let $W_i = \varepsilon_i X_i^{\sharp(r)}$ and $\nu = 4r+4z$. By $\tilde{X}_i\varepsilon_i = W_i + X_i^{\flat(r)}\varepsilon_i$, $\mathrm{E}[W_i] = 0$, and expanding the corresponding quadratics,

$$\frac{1}{n}\mathrm{var}(\sum_{i\in S}\varepsilon_i X_i^{\sharp(r)}) - \mathrm{E}[\widehat{\Omega}_0^K] = \frac{1}{n}\mathrm{E}[\sum_{i\in S^2}(1-K_{i_1i_2})W_{i_1}W_{i_2} + \sum_{i\in S^2}(-K_{i_1i_2})X_{i_1}^{\flat(r)}\varepsilon_{i_1}(2W_{i_2} + X_{i_2}^{\flat(r)}\varepsilon_{i_2})].$$

Use $T = S$ and Δ defined with $x \ge 1$ as in Lemma 1. Then $|\Delta| \le nL_{\mathrm{growth}}^{\log x+2}$ and for $i \in \Delta$, using Lipshitz part of the Kernel Regularity Assumption, $|1-K_{i_1i_2}| \le L_{\mathrm{kern}}(x/h)$. Note also that, as $\mathrm{E}[W_{i_1}] = 0$, $|\mathrm{E}[W_{i_1}W_{i_2}]| \le \exp(-(x-(4r+4z))/L_{\mathrm{mix}})(\mathrm{E}[|W_{i_1}W_{i_2}| + \mathrm{E}[|W_{i_1}|]\mathrm{E}[|W_{i_2}|]) \le \exp(-(x-\nu)/L_{\mathrm{mix}})2L_{\mathrm{mom}}^{1/2}f_{\mathrm{PAR}}(r,x)^{1/2}$. Then $n^{-1}\Big(\sum_{i\in\Delta}|1-K_{i_1i_2}| + \sum_{i\in S^2\setminus\Delta}2\exp(-(x-\nu)/L_{\mathrm{mix}})\Big)L_{\mathrm{mom}}^{1/2}f_{\mathrm{PAR}}(r,x)^{1/2}$ $\le \Big(L_{\mathrm{kern}}L_{\mathrm{growth}}^{\log x+2}(x/h) + 2n\exp(-(x-\nu)/L_{\mathrm{mix}})\Big)L_{\mathrm{mom}}^{1/2}f_{\mathrm{PAR}}(r,x)^{1/2}$. Additionally, by Hölder inequality and kernel having $|K_{i_1i_2}| \le 1$, $|K_{i_1i_2}\mathrm{E}[X_{i_1}^{\flat(r)}\varepsilon_{i_1}(2W_{i_2}+X_{i_2}^{\flat(r)}\varepsilon_{i_2})]| \le 2L_{\mathrm{basis}}f_{\mathrm{TRNC}}(r,x)^{1/4}L_{\mathrm{mom}}^{1/2}f_{\mathrm{PAR}}(r,x)^{1/4} + L_{\mathrm{mom}}^{1/2}L_{\mathrm{basis}}^{2}f_{\mathrm{TRNC}}(r,x)^{1/2}$. Applying $n^{-1}\sum_{i\in S^2}$ to previous bound and adding gives $|\frac{1}{n}\mathrm{var}(\sum_{i\in S}\varepsilon_i X_i^{\sharp(r)}) - \mathrm{E}[\widehat{\Omega}_0^K]| \le f_{\mathrm{KER}}(r,x)$. ■

Assembly of lemmas 1–15. Theorem 1 now follows by assembling the lemmas. For any data generating process in $\mathcal{P}_F$ and tuning parameters in $\mathcal{T}_F$ let

$$\epsilon_0 = |\Pr(\beta_0 \in \widehat{C}) - (1-\alpha)|.$$

Suppose that $f_{\mathrm{COND}} > 0$, or equivalently, $b > L_{\mathrm{basis}}^4$. Let

$$\begin{aligned}
r_* &= z^{1/2},\\
\nu_* &= 4r_* + 4z\\
x_{1*} &= \log(n)^2,\ x_{2*} = \nu_* + \log(n)^2,\\
u_* &= f_{\mathrm{LLN}}(S, x_{2*}, \nu_*)^{1/4}\\
\rho_* &= f_{\mathrm{COND}}^{-3/2}(L_{\mathrm{mom}} f_{\mathrm{PAR}}(r_*, x_{1*}))^{3/4} \max_{\varnothing \neq T \subseteq S} f_{4\mathrm{TH}}(T, x_{2*}, \nu_*)^{3/4},\\
y_* &= n^{1/(4\dim(S))},\\
\upsilon_* &= 4\sqrt{\log n} f_{\mathrm{COND}}^{-1} \max(f_{\mathrm{VAR}}(r_*, x_{2*})^{1/4}, f_{\mathrm{RES}}(r_*, x_{2*}, u_*)^{1/2}, 4 f_{\mathrm{KER}}(r_*, x_{2*})).
\end{aligned}$$

Let $\widehat{\phi} = \widehat{V}^{-1/2}(\widehat{\beta} - \beta_0)$ where $\widehat{V} = n^{-1}\widehat{\Omega}^{\kappa}/\widehat{Q}^2$ and $\widehat{Q} = n^{-1}\sum_{i\in S}\tilde{X}_i^2$. Let $Q_0 = \mathrm{E}[\widehat{Q}]$. Set $\Xi = \sum_{i\in S} X_i^{\sharp(r_*)}\varepsilon_i$. and $\sigma = \mathrm{var}(\Xi)^{1/2}$. Let $\phi_0 = \sigma^{-1}\Xi$. Note that $\mathrm{E}[\phi_0] = 0$ and $\mathrm{var}(\phi_0) = 1$. Let $\Gamma = \sum_{i\in S} X_i^{\flat(r_*)}\varepsilon_i$. Additionally $\widehat{\phi} = n^{-1/2}(\Xi+\Gamma)(\widehat{\Omega}^{\kappa})^{-1/2}$. Let Φ denote the standard Gaussian cumulative distribution function. Let $t_\alpha = q_{1-\alpha/2}$ and recall that $\beta_0 \in \widehat{C}$ precisely when $|\widehat{\phi}| \le t_\alpha$. Next derive a bound assuming $f_{\mathrm{COND}} > 0$. This will be accounted for later.

Comparing to a Gaussian,

$$\begin{aligned}
\epsilon_0 &\le 2\sup_{t\in\mathbb{R}} |\Pr(\phi_0 \le t) - \Phi(t)| + \sup_{t\in\mathbb{R}} \Pr(|\phi_0 - t| < \upsilon_*) + \Pr(|\widehat{\phi} - \phi_0| > \upsilon_*)\\
&\le 4\sup_{t\in\mathbb{R}} |\Pr(\phi_0 \le t) - \Phi(t)| + \sup_{t\in\mathbb{R}} \Pr(|N(0,1) - t| < \upsilon_*) + \Pr(|\widehat{\phi} - \phi_0| > \upsilon_*).
\end{aligned}$$

Note that $\Pr(|N(0,1) - t| < \upsilon_*) \le \sqrt{2/\pi}\,\upsilon_*$. Lemma 6 is applicable to Ξ with choices of delay ν_*, truncation $x = y_*$ moment bound $\rho = \rho_*$. Note that the reason that the denominator in the definition of ρ_* can be used to justify application of Lemma 6 is that Lemma 10 derives $\mathrm{E}[(|T|^{-1/2}\sum_{i\in T} X_i^{\sharp(r)}\varepsilon_i)^2] \ge f_{\mathrm{COND}}$. Similarly, the numerator defining ρ_* is applicable as a bound for corresponding third moment quantities for Ξ. Then Lemma 6 gives

$$\begin{aligned}
\sup_{t\in\mathbb{R}} \big|\Pr(\phi_0 \le t) - \Phi(t)\big| \le{}& 15.12\, n^{-1/2}\rho_* L_{\mathrm{growth}}^{\log y_* + 3/y_* + 5/2} + 6n\exp(-(y_* - \nu_*)/L_{\mathrm{mix}})\\
&+ 16\rho_*^{4/9}\big(y_*^{-1/3} + n^{2/3}\exp(-(y_* - \nu_*)/(3L_{\mathrm{mix}}))\big) L_{\mathrm{growth}}^{2/y_* + 2/3}.
\end{aligned}$$

Next, recall $\widehat{\phi} = n^{-1/2}(\Xi + \Gamma)(\widehat{\Omega}^{\kappa})^{-1/2}$ and $\phi_0 = \Xi\sigma^{-1}$.

$$|\widehat{\phi} - \phi_0| \le n^{-1/2}|\Gamma|(\widehat{\Omega}^{\kappa})^{-1/2} + n^{-1/2}|\Xi||(\widehat{\Omega}^{\kappa})^{-1/2} - (\sigma^2/n)^{-1/2}|.$$

For bounding $\Pr(|\widehat{\phi} - \phi_0| > \upsilon_*)$, let $\delta_\Omega = |\widehat{\Omega}^{\kappa} - \sigma^2/n|$. Define the event $\mathcal{E} = \{\delta_\Omega \le (\sigma^2/n)/2\}$. Use $|a^{-1/2} - b^{-1/2}| \le \frac{|a-b|}{2\min(a,b)^{3/2}}$. Then on $\mathcal{E}$, $(\widehat{\Omega}^{\kappa}) \ge (\sigma^2/n)/2$ so $(\widehat{\Omega}^{\kappa})^{-1/2} \le \sqrt{2}(\sigma^2/n)^{-1/2}$ and

$$|\widehat{\phi} - \phi_0| \le \Big|\frac{\Xi}{\sqrt{n}}\Big|\frac{|\sigma^2/n - \widehat{\Omega}^{\kappa}|}{2((\sigma^2/n)/2)^{3/2}} + \frac{\sqrt{2}|\Gamma|}{\sqrt{n}(\sigma^2/n)^{1/2}} = |\phi_0|\frac{\sqrt{2}\delta_\Omega}{\sigma^2/n} + \frac{\sqrt{2}|\Gamma|}{\sqrt{n}(\sigma^2/n)^{1/2}}.$$

Also, by the same reason as in the discussion on using ρ_* for applying Lemma 6, $\sigma^2/n \ge f_{\mathrm{COND}}$. Then note that there is the inclusion of events $\{|\widehat{\phi} - \phi_0| > \upsilon_*\} \subseteq \mathcal{E}^c \cup \{|\phi_0|\sqrt{2}\delta_\Omega/f_{\mathrm{COND}} > \upsilon_*/2\} \cup \{\sqrt{2}|\Gamma|/(\sqrt{n}f_{\mathrm{COND}}^{1/2}) > \upsilon_*/2\}$. The middle event can also decomposed $\{|\phi_0|\sqrt{2}\delta_\Omega/f_{\mathrm{COND}} > \upsilon_*/2\} \subseteq \{|\phi_0| > \sqrt{2\log n}\} \cup (\{|\phi_0|\sqrt{2}\delta_\Omega/f_{\mathrm{COND}} > \upsilon_*/2\} \cap \{|\phi_0| \le \sqrt{2\log n}\})$. This leads to

$$\begin{aligned}
\Pr(|\widehat{\phi} - \phi_0| \ge \upsilon_*) \le{}& \Pr\Big(\delta_\Omega > \frac{f_{\mathrm{COND}}}{2}\Big) + \Pr(|\phi_0| > \sqrt{2\log n}) + \Pr\Big(\delta_\Omega > \frac{\upsilon_* f_{\mathrm{COND}}}{4\sqrt{\log n}}\Big)\\
&+ \Pr\Big(|\Gamma| > \frac{\upsilon_* n^{1/2} f_{\mathrm{COND}}^{1/2}}{2\sqrt{2}}\Big).
\end{aligned}$$

Note that $\Pr(|\phi_0| \geq \sqrt{2\log n}) \leq 2(1-\Phi(\sqrt{2\log n})) + 2\sup_{t\in\mathbb{R}}|\Pr(\phi_0 \leq t) - \Phi(t)|$, to which Lemma 6 is applicable. For $|\Gamma|$ note Lemmas 8 and 9 give a bound $\mathrm{E}[\Gamma^4] \leq n^4 \max_{i\in S}\mathrm{E}[(X_i^{\flat(r_*)}\varepsilon_i)^4] \leq n^4 L_{\mathrm{mom}} L_{\mathrm{basis}}^4 f_{\mathrm{TRNC}}(r_*, x_{1*})$. By Markov's inequality, it holds that $\Pr(|\Gamma| > \upsilon_* n^{1/2} f_{\mathrm{COND}}^{1/2}/(2\sqrt{2})) \leq 64 n^2 L_{\mathrm{mom}} L_{\mathrm{basis}}^4 f_{\mathrm{TRNC}}(r_*, x_{1*})/(\upsilon_*^4 f_{\mathrm{COND}}^2)$. Finally, $\Pr(|\delta_\beta| \geq u_*) \leq 4u_*^{-2} f_{\mathrm{COND}}^{-2} f_{\mathrm{BETA}}(r_*, x_{2*}, u_*)$ and note that $\{|\widehat{\Omega}_0^K - \widehat{\Omega}^K| > c/3\} \subseteq \{|\delta_\beta| \geq u_*\} \cup \bigcup_{j\leq 5}\{|\delta_j(u_*)| > (c/3)/5\}$ and on $\{|\delta_\beta| \leq u_*|\}$, $|\delta_j(\delta_\beta)| \leq |\delta_j(u_*)|$.

Let $c > 0$. Then

$$\begin{aligned}
\Pr(\delta_\Omega > c) &\leq 1\{|\sigma^2/n - \mathrm{E}[\widehat{\Omega}_0^K]| > c/3\} + \Pr(|\mathrm{E}[\widehat{\Omega}_0^K] - \widehat{\Omega}_0^K| > c/3) + \Pr(|\widehat{\Omega}_0^K - \widehat{\Omega}^K| > c/3). \\
&\leq 1\{f_{\mathrm{KER}}(r_*, x_{2*}) > \frac{c}{3}\} + \frac{9}{c^2} f_{\mathrm{VAR}}(r_*, x_{2*}) + \Pr(|\delta_\beta| > u_*) + \sum_{j=1}^{5} \Pr(|\delta_j(u_*)| \geq \frac{c}{3}/5) \\
&\leq 1\{f_{\mathrm{KER}}(r_*, x_{2*}) > \frac{c}{3}\} + \frac{9}{c^2} f_{\mathrm{VAR}}(r_*, x_{2*}) + \frac{4 f_{\mathrm{BETA}}(r_*, x_{2*}, u_*)}{u_*^2 f_{\mathrm{COND}}^2} + \frac{75}{c} f_{\mathrm{RES}}(r_*, x_{2*}, u_*).
\end{aligned}$$

Applying this with $c = \frac{f_{\mathrm{COND}}}{2}$ and $c = \frac{\upsilon_* f_{\mathrm{COND}}}{4\sqrt{\log n}} = \max(f_{\mathrm{VAR}}(r_*, x_{2*})^{1/4}, f_{\mathrm{RES}}(r_*, x_{2*}, u_*)^{1/2}, 4 f_{\mathrm{KER}}(r_*, x_{2*}))$ and recalling that $f_{\mathrm{CLT}}(n, \rho_*, y_*, \nu_*) = 15.12\, n^{-1/2}\rho_* L_{\mathrm{growth}}^{\log y_* + 3/y_* + 5/2} + 6n\exp(-(y_* - \nu_*)/L_{\mathrm{mix}}) + 16\rho_*^{4/9}\big(y_*^{-1/3} + n^{2/3}\exp(-(y_* - \nu_*)/(3L_{\mathrm{mix}}))\big) L_{\mathrm{growth}}^{2/y_* + 2/3}$ gives

$$\begin{aligned}
\epsilon_0 &\leq 6 f_{\mathrm{CLT}}(n, \rho_*, y_*, \nu_*) \\
&\quad + \sqrt{2/\pi}\Big(\frac{4\sqrt{\log n}\max(f_{\mathrm{VAR}}(r_*, x_{2*})^{1/4}, f_{\mathrm{RES}}(r_*, x_{2*}, u_*)^{1/2}, 4 f_{\mathrm{KER}}(r_*, x_{2*}))}{f_{\mathrm{COND}}}\Big) \\
&\quad + 2(1 - \Phi(\sqrt{2\log n})) \\
&\quad + 64 n^2 L_{\mathrm{mom}} L_{\mathrm{basis}}^4 f_{\mathrm{TRNC}}(r_*, x_{1*})/(\upsilon_*^4 f_{\mathrm{COND}}^2) \\
&\quad + 1\{f_{\mathrm{KER}}(r_*, x_{2*}) \not< \frac{f_{\mathrm{COND}}}{6}\} + \frac{36 f_{\mathrm{VAR}}(r_*, x_{2*})}{f_{\mathrm{COND}}^2} + \frac{4 f_{\mathrm{BETA}}(r_*, x_{2*}, u_*)}{u_*^2 f_{\mathrm{COND}}^2} + \frac{150 f_{\mathrm{RES}}(r_*, x_{2*}, u_*)}{f_{\mathrm{COND}}} \\
&\quad + 9 f_{\mathrm{VAR}}(r_*, x_{2*})^{1/2} + \frac{4 f_{\mathrm{BETA}}(r_*, x_{2*}, u_*)}{u_*^2 f_{\mathrm{COND}}^2} + 75 f_{\mathrm{RES}}(r_*, x_{2*}, u_*)^{1/2}.
\end{aligned}$$

Let $f_F(n, h, z, b)$ be defined as this final expression, and thus $\epsilon_0 \leq f_F(n, h, z, b)$. For every fixed frame, a sufficient condition for $f_F(n, h, z, b) \to 0$ is $n \to \infty$ and

$$\begin{aligned}
&\frac{z}{\log(n)^4} \to \infty, && \frac{n}{z\,\text{max-vol}(z)\text{max-vol}(h)} \to \infty, \\
&\frac{h}{z^2\,\text{max-vol}(z)} \to \infty, && \frac{n}{\text{max-vol}(z)^{8\dim(S)}} \to \infty.
\end{aligned}$$

■

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 Σ. So we generate variables X_i and Y_i with:

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

Where X has components X_i and Y has components Y_i. Σ has variances of one and off-diagonal elements (i, j) given by $\exp(-d_{ij}^{\text{Euc}}/\theta)$ with d_{ij}^{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 Σ has parameter $\theta = \sqrt{2}/10$. To better understand the 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 G_i terms and different bandwidths using a Gaussian kernel, with variance $\sigma^2 I$. [6] The bandwidth is described by headings .05,.10, .15 which give the value of 2σ for each kernel. The second panel uses the same HAC estimator but adds an 8×8 tensor product of triangular B-splines serving as G_i to the regression.[7]

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 values of ρ. 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 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 ρ approaches one the variance of sample means is similar to its analog for a highly serially correlated AR1 process.

[6] 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.

[7] 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.

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 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 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.

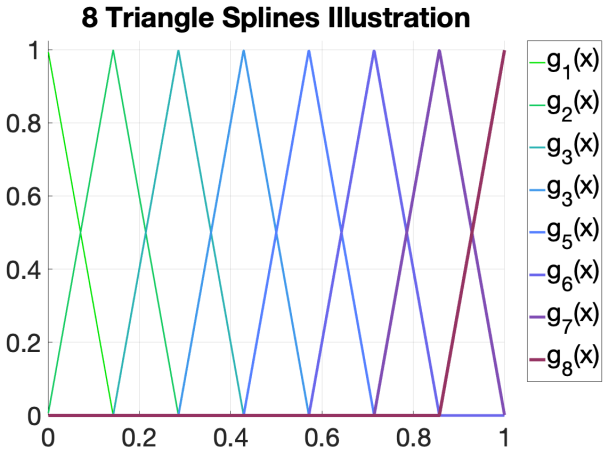

Figure 1: The figure illustrates our set of eight triangle splines in each individual coordinate dimension. Each is zero for all coordinates outside the base of its triangle. Our tensor spline is comprised of all products of the eight vertical and eight horizontal coordinate splines.

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 differing values of ρ 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 ρ are also about .19 and then slowly increase as ρ increases until about .20 at $\rho = .8$. HAC coverage probabilities remain 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

		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 labeled '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: 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.

$\rho = .9, 1$ 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 our DGP 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 $\widehat{\beta}$ estimator but it will also remove some of the identifying variation in X_i 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, 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

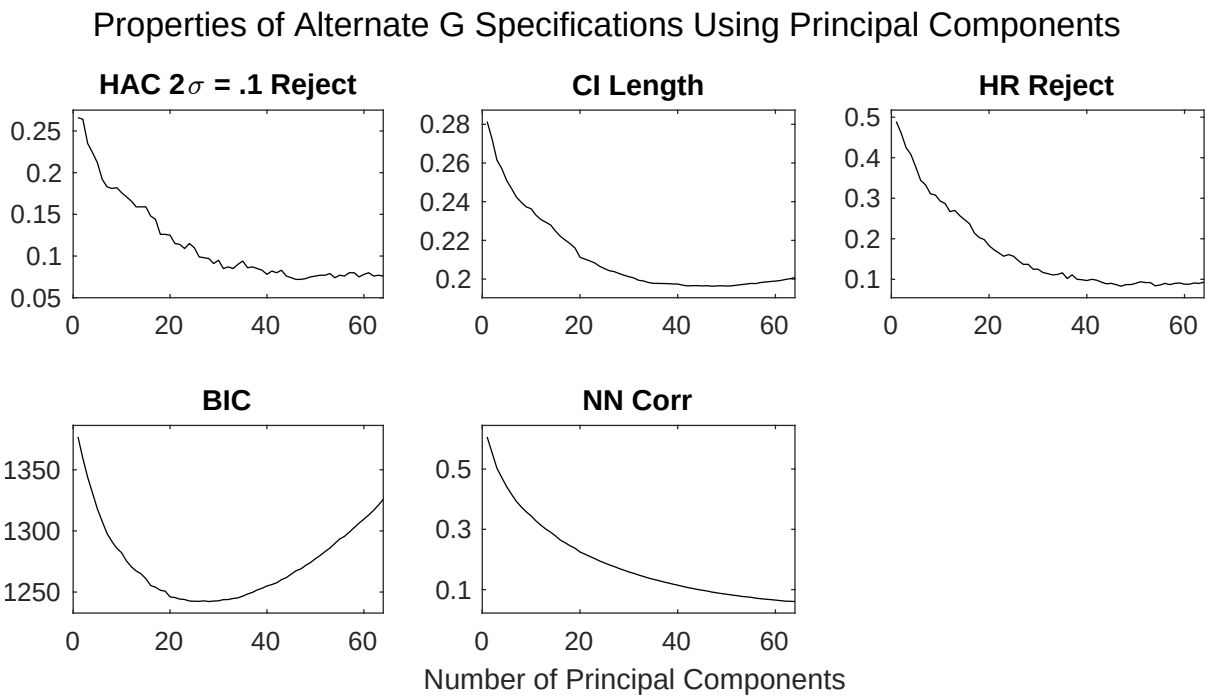

Figure 2: The horizontal axis indexes the number of principal components from an 8 by 8 tensor product of triangle B-splines used in the spatial basis, G_i. The DGP uses $\rho = .8$.

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 bandwith 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 it's 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 these simulations, we examine tensor products of 10 triangle B-splines in each dimension to form the set of candidate G_i. [8]

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

[8] Results for 8 by 8 tensors are qualitatively similar and available upon request.

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 ρ but not drastically. For lower levels of ρ 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 bandwith 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.[9] The procedure for each simulation is: (1) fit the approximate model to the simulated dataset to get an approximate DGP: $F(\widehat{\tau})$. (2) Using the real locations, generate a large sample of Monte Carlo (MC) draws from $F(\widehat{\tau})$. (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. [10]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.

Table 5 presents simulation results for rejection frequencies and 95% confidence interval length for the CHKV.[11] This Table shows a very different pattern in rejection frequencies as ρ varies compared with Tables 1 and 3. There are minimal size distortions for high and low ρ but distortions of up to 8% for medium values. Because the CHKV method uses simulations from an approximate model $F(\tau)$ to

[9]This grid contained all combinations of Gaussian Kernel bandwidth (2σ) 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].

[10]We calculated power for 2 alternatives, slope = $\pm 3/N^{1/2}$

[11]The rows of the Table refer to differing values of ρ in our simulation DGP, just as in previous tables. Likewise the second column provides the true DGP correlation at distance of .1. Rej. 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.

Simulation Results: Rejection Frequencies

		No Splines				Triangle Splines					
		HAC				HAC					
ρ	Corr	.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 (ρ). 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 true correlation at distance $h = 0.1$. 1000 simulations.

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 ρ 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 ρ than for medium ρ. 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 having a high quality model for $F(\tau)$ is a drawback.

Overall, we think the choice of NN versus CHKV boils down to whether the researcher feels they have a good guess at an approximating model $F(\tau)$, using CHKV if so and NN if not. In the empirical example in the next section we use NN.

6. Empirical Illustration Revisiting the Relation Between Government Centralization Measures and Night-Time Illumination Measures

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.

Simulation Results: Confidence Interval Lengths

		HAC			
ρ	Corr	.05	.10	.15	HR
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 **10x10** tensor of triangular B-splines that minimize the residuals' absolute nearest neighbor correlation. 1000 simulations. Column *Corr* shows the true correlation at distance $h = 0.1$. 1000 simulations.

Simulation Results: Bandwith/Basis Selection Method

ρ	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.

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 observation locations are plotted in Figure 4 below.

Figure 3 illustrates our NN (nearest neighbor) residual absolute correlation objective function for different size tensors, from 9 by 9 to 12 by 12. The 12 by 12 tensor yields an NN residual absolute 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 focus on 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 presents point estimates and standard errors of the treatment slope of the original MP regression and our estimates for G_i defined using different numbers of PCs and alternate HAC estimators.[12] Here, we differ from our kernel choice in simulation sections and use a uniform kernel for our HAC estimators for ease in interpretation. In Table 6 Panel A, 65 PCs yield a NN residual correlation of approximately zero while, in Table 6 Panel B, 50 PCs result 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 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.

Government Centralization and Night-Time Illumination

	Original	HAC(150)	HAC(250)	HAC(350)
Panel A: 65 PCs				
Point Estimate	.41	.23	.23	.23
Std Error	.12	.07	.06	.07
Panel B: 50 PCs				
Point Estimate	.41	.22	.22	.22
Std Error	.12	.07	.07	.07

Table 6: Point Estimates and Standard Errors, Original and HAC with uniform kernel bandwidths of 150,250, and 350km. G is 65 PCs from 12 by 12 tensor product of triangle B-Splines in Panel A. G is 50 PCs from 12 by 12 tensor product of triangle B-splines in Panel B.

[12] To illustrate an estimated $G\widehat{\gamma}$ Figure 5 displays an estimated "surface" for one G_i choice.

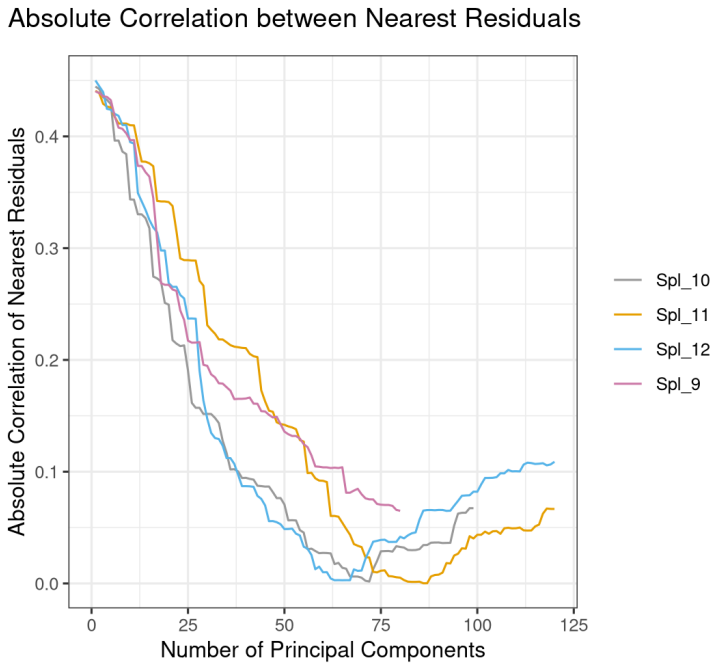

Figure 3: Absolute NN correlations for choices of tensor from 9 by 9 to 12 by 12 and alternate numbers of PCs.

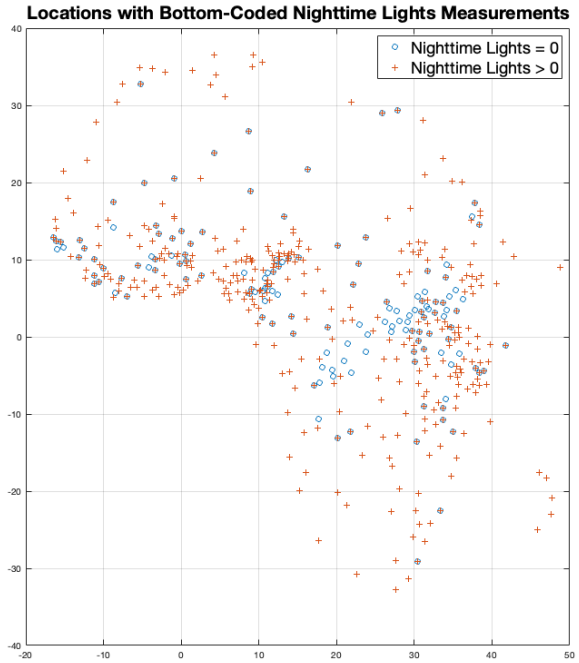

Figure 4: The Figure illustrates locations from the data used in [Michalopoulos and Papaioannou, 2013] in which nighttime lights measurements are sufficiently low to be bottom coded. Coordinates are longitude and latitude. Circles are used for bottom-coded observations, plus signs for the remainder.

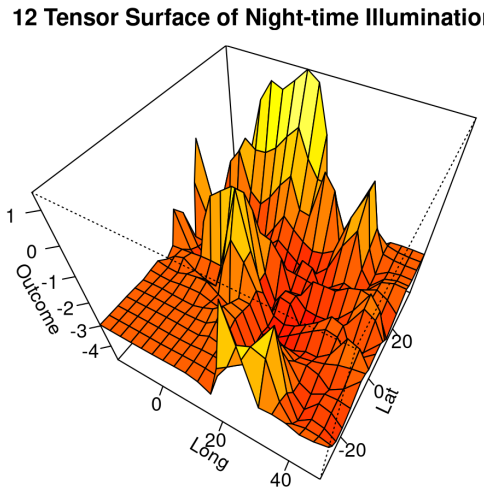

Figure 5: The Figure illustrates estimates $G\widehat{\gamma}$ versus coordinates as an estimated "surface" using 60 PCs from a 12 by 12 tensor product of triangle B-splines.

7. Conclusion and Discussion

In this paper, we have presented a method that makes it easier to conduct spatial dependence robust inference using spatially indexed data. The general structure of the method has two steps. First, the method augments the regression specification with spatially localized regressor terms that we call spatial basis terms. These spatial basis terms have true coefficients of zero but in small samples absorb some of the spatial dependence in residuals and scores. The augmented specification has the advantage of having less severe spatial correlation of scores making inference easier. Spatial dependence in scores is reduced but not eliminated, so the second component of the method is application of a HAC method to obtain spatial dependence robust confidence intervals. The requisite tuning parameters to implement HAC inference are much easier to choose with the reduced-correlation scores from our spatial basis augmented first stage, compared with the original scores.

The paper derives statistical properties of our procedure and gives a precise sense in which our confidence sets achieve approximately (i.e., asymptotically) nominal coverage. We present formal restrictions on the choice of spatial basis terms and HAC tuning parameters to ensure consistency of point estimates and validity of confidence intervals. The simulation study we present shows that our approximations work well in datasets as small as $n = 500$, with various nontrivial levels of spatial correlation; see Section 4 above. It also demonstrates that basis choice via minimizing a nearest neighbor absolute correlation in residuals, while ad hoc, works well even with high levels of spatial dependence. While our analysis focuses on using a second stage HAC estimator, augmenting regressions with our spatial basis terms is feasible with many spatial dependence robust inference procedures.

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*.

[Gonçalves and Ng, 2024] Gonçalves, S. and Ng, S. (2024). Imputation of counterfactual outcomes when the errors are predictable. *Journal of Business & Economic Statistics*, 42(4):1107–1122.

[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.

[Mendel and Naor, 2006] Mendel, M. and Naor, A. (2006). Ramsey partitions and proximity data structures. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 109–118.

[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.