Nonparametric Spatial Modeling towards the Mode

Tao Wang[a] Weixin Yao[b]

a.University of Victoria b.University of California Riverside

Supplementary Material

The Supplementary Material contains comments for MEM algorithm and theoretical conditions, boundary analysis, additional simulations, generalizations to additive and extended spatial modal regression models, as well as technical proofs of the main theorems and supporting lemmas.

S1 Comments for MEM Algorithm 1

This section provides theoretical and practical insights into the MEM algorithm used in our spatial modal regression framework. Particularly, we examine numerical stability, convergence properties, and offer theoretical guarantees comparing MEM with mean-based estimators.

S1-1 Numerical Stability

A potential numerical challenge in the implementation of MEM Algorithm 1 arises in the M-step, which involves the inversion of the weighted local

design matrix ($\mathbf{X}^{*T}W_{\mathbf{X}}\mathbf{X}^{*}$). The weight matrix $W_{\mathbf{X}}$ is diagonal with entries $\pi(\mathbf{i} \mid \boldsymbol{\theta}^{(g)})$, defined by

$$\pi(\mathbf{i} \mid \boldsymbol{\theta}^{(g)}) \propto \phi\left(\frac{Y_{\mathbf{i}} - a^{(g)} - \mathbf{b}^{(g)T}(\mathbf{X_i} - \mathbf{x})}{h_1}\right) K\left(\frac{\mathbf{X_i} - \mathbf{x}}{h_2}\right),$$

where both $\phi(\cdot)$ and $K(\cdot)$ are Gaussian kernels with unbounded support and exponentially decaying tails. From a theoretical perspective, as the bandwidth $h_2 \to 0$, the kernel weights $K((\mathbf{X_i}-\mathbf{x})/h_2)$ tend towards zero for observations distant from the evaluation point $\mathbf{x}$. Consequently, $\pi(\mathbf{i} \mid \boldsymbol{\theta}^{(g)})$ becomes highly sparse, and the effective number of influential observations may become insufficient to ensure that the weighted local design matrix $\mathbf{X}^{*T}W_{\mathbf{X}}\mathbf{X}^{*}$ remains well-conditioned. In such cases, the matrix inversion can suffer from high condition numbers, introducing substantial numerical error due to finite precision arithmetic.

This issue is especially pronounced in high-dimensional settings or in regions with sparse covariate support. If the design matrix has a high condition number, numerical errors can enter and propagate through the estimation steps, which can severely degrade the performance of the algorithm. This sensitivity to h_2 and the resulting weight sparsity reflects a fundamental trade-off in kernel smoothing, i.e., improving local fidelity while risking numerical instability. In the paper, this theoretical risk is effectively mitigated by the bandwidth selection procedure developed in Subsection

2.4. The proposed selection is based on minimizing the asymptotic mean squared error (AsyMSE) of the local modal estimator $\hat{m}(\mathbf{x})$, which balances squared bias and variance

$$\text{AsyMSE}(\hat{m}(\mathbf{x})) = \text{Bias}^2(\hat{m}(\mathbf{x})) + \text{Var}(\hat{m}(\mathbf{x})).$$

The resulting optimal bandwidth $\hat{h}_2$ satisfies the rate $\tilde{\mathbf{n}}^{-\frac{1}{d+7}}$, which corresponds to more smoothing than the classical mean regression rate. This is because modal estimation involves greater curvature sensitivity, leading to higher variance near the density peak; hence, a slightly larger bandwidth is needed to ensure stability. This optimal rate ensures that while the estimator remains sufficiently localized to capture modal features, it retains a large enough effective sample size to avoid the degeneracy associated with vanishing weights and poor matrix conditioning. Consequently, the AsyMSE-optimal bandwidth plays a dual role, i.e., it governs the local approximation error and simultaneously acts as an implicit regularization device that stabilizes the M-step by preserving invertibility of the weighted local design matrix. For additional robustness in finite samples, particularly when multicollinearity or sparse data regions are present, one may

Under the optimal bandwidth regime $h_1 \asymp h_2 \asymp \tilde{\mathbf{n}}^{-1/(d+7)}$, the effective sample size contributing to the MEM estimator is of order $\tilde{\mathbf{n}} h_2^d h_1^3 \asymp \tilde{\mathbf{n}}^{1-\frac{d+3}{d+7}} = \tilde{\mathbf{n}}^{\frac{4}{d+7}}$, which diverges as $\tilde{\mathbf{n}} \to \infty$. This ensures that, asymptotically, the weighted local design matrix remains well-conditioned with high probability, and the numerical instability due to kernel weight sparsity is avoided.

supplement this procedure with an explicit regularization strategy. Specifically, the weighted matrix inversion can be replaced with a ridge-adjusted version $(\mathbf{X}^{*T}W_{\mathbf{X}}\mathbf{X}^{*}+\lambda I)^{-1}$ for a small regularization parameter $\lambda > 0$ (e.g., $\lambda = 10^{-6} \times tr(\mathbf{X}^{*T}W_{\mathbf{X}}\mathbf{X}^{*})$), thereby ensuring invertibility.

S1-2 Convergence Behavior and Global Optimality

The MEM algorithm plays a central role in our spatial modal regression framework. However, a central concern lies in the fact that MEM algorithm, like many EM-type algorithms, is not guaranteed to converge globally. This is particularly relevant because the primary estimand in modal regression is the mode, not merely a local maximizer of the objective function. To address this concern, we provide a detailed theoretical analysis of the algorithm's convergence behavior and propose practical heuristics to ensure robust performance in empirical applications.

We first formalize the convergence properties under regularity conditions. The result mirrors the Newton-Kantorovich framework (Ortega and Rheinboldt, 1970) but is adapted to our modal objective.

Theorem S1 (Local Quadratic Convergence). *Let $\hat{\boldsymbol{\theta}}$ denote a strict local maximizer of the smoothed kernel objective $Q_{\tilde{\mathbf{n}}}(\boldsymbol{\theta})$ defined in (2.3). Suppose (i) $Q_{\tilde{\mathbf{n}}}(\boldsymbol{\theta})$ is three-times continuously differentiable in a neighborhood of $\hat{\boldsymbol{\theta}}$;*

(ii) the gradient $\nabla Q_{\tilde{\mathbf{n}}}(\hat{\boldsymbol{\theta}}) = 0$ *and Hessian* $\nabla^2 Q_{\tilde{\mathbf{n}}}(\hat{\boldsymbol{\theta}})$ *is negative definite; (iii) the MEM update map* $T(\boldsymbol{\theta})$ *satisfies a second-order expansion, i.e.,* $T(\boldsymbol{\theta}) = \hat{\boldsymbol{\theta}} + \frac{1}{2}H(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2 + o(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2)$*, where* H *is a symmetric matrix derived from curvature of* $Q_{\tilde{\mathbf{n}}}$*; and (iv) the initial point* $\boldsymbol{\theta}^{(0)}$ *lies sufficiently close to* $\hat{\boldsymbol{\theta}}$*. Then, the sequence* $\{\boldsymbol{\theta}^{(g)}\}$ *generated by MEM satisfies*

$$\|\boldsymbol{\theta}^{(g+1)} - \hat{\boldsymbol{\theta}}\| \leq C\|\boldsymbol{\theta}^{(g)} - \hat{\boldsymbol{\theta}}\|^2,$$

for some constant $C > 0$*, i.e., MEM converges quadratically.*

Proof. Let $T(\boldsymbol{\theta})$ be the MEM update map. By Taylor expansion around $\hat{\boldsymbol{\theta}}$ and using $T(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\theta}}$, we obtain

$$T(\boldsymbol{\theta}) = \hat{\boldsymbol{\theta}} + J(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T H_T(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}),$$

where $J = \nabla T(\hat{\boldsymbol{\theta}})$, and $H_T(\tilde{\boldsymbol{\theta}})$ is a third-order tensor evaluated at some $\tilde{\boldsymbol{\theta}}$ on the line segment between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$. Because $T(\boldsymbol{\theta}^{(g+1)}) = \boldsymbol{\theta}^{(g+1)}$, the expansion simplifies to

$$\boldsymbol{\theta}^{(g+1)} - \hat{\boldsymbol{\theta}} = J(\boldsymbol{\theta}^{(g)} - \hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta}^{(g)} - \hat{\boldsymbol{\theta}})^T H_T(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta}^{(g)} - \hat{\boldsymbol{\theta}}).$$

Taking norms and applying the triangle inequality, we can obtain

$$\|\boldsymbol{\theta}^{(g+1)} - \hat{\boldsymbol{\theta}}\| \leq \|J\| \cdot \|\boldsymbol{\theta}^{(g)} - \hat{\boldsymbol{\theta}}\| + C_1\|\boldsymbol{\theta}^{(g)} - \hat{\boldsymbol{\theta}}\|^2,$$

for some $C_1 > 0$ depending on the local curvature. By assumption listed in Theorem S1, $\|J\| < 1$, so there exists $\delta > 0$ such that for all $\|\boldsymbol{\theta}^{(g)} - \hat{\boldsymbol{\theta}}\| < \delta$, we have $\|\boldsymbol{\theta}^{(g+1)} - \hat{\boldsymbol{\theta}}\| \leq C\|\boldsymbol{\theta}^{(g)} - \hat{\boldsymbol{\theta}}\|^2$, for some $C > 0$ after absorbing the linear term into the quadratic. This completes the proof. □

Theorem S1 shows that the MEM algorithm exhibits local quadratic convergence under standard smoothness and identifiability conditions, mirroring the behavior of Newton-type methods near a local mode. This guarantees rapid convergence once the iterates enter a sufficiently small neighborhood of a strict local maximum. However, because the kernel-smoothed objective may be non-convex and possess multiple local optima, the MEM algorithm is not guaranteed to converge to the global mode unless appropriately initialized. This limitation is inherent to EM-type methods and is particularly important here, given that the estimand of interest is the mode, not just any local maximizer. To address this issue in practice, we propose three complementary heuristics: (i) initialization from multiple starting points, such as using local linear mean or quantile regression estimators as pilot values. This increases the chance of exploring different basins of attraction in the objective landscape; (ii) employing tempered EM variants or injecting controlled stochastic perturbations during early iterations. Tempered EM begins with a large bandwidth h_1, resulting in a smoother likelihood surface, and gradually anneals to the target bandwidth, helping the algorithm avoid convergence to suboptimal modes. Similarly, mild stochastic noise can help navigate past flat or shallow regions of the objective; (iii) selecting the final estimator by comparing the kernel-smoothed

likelihood values across candidate solutions and retaining the one that attains the highest value. These heuristics are straightforward to implement and collectively improve the algorithm's robustness to poor initializations. Section 4 in the paper provides detailed simulation studies illustrating that, across various error structures and sample sizes, the MEM algorithm consistently recovers the global mode.

To further address concerns regarding the practical and theoretical value of our proposed MEM estimator, we establish a formal result comparing the kernel-smoothed conditional likelihood achieved by MEM to that of standard local linear mean regression. In particular, we show that under non-Gaussian error distributions, especially those exhibiting asymmetry or multimodality, the MEM estimator yields a strictly higher value of the kernel-based objective function. This result offers theoretical support for the use of modal regression in such settings, as the MEM estimator more accurately targets the mode of the conditional distribution rather than the mean, which may be misleading or nonexistent. The following theorem quantifies this advantage.

Theorem S2 (Relative Likelihood Superiority of MEM). *Let* $(Y_{\mathbf{i}}, \mathbf{X_i})_{\mathbf{i}\in\mathbf{I}_{\tilde{\mathbf{n}}}}$ *follow the model* $Y_{\mathbf{i}} = m(\mathbf{X_i}) + \varepsilon_{\mathbf{i}}$ *with* $Mode(\varepsilon_{\mathbf{i}} \mid \mathbf{X_i}) = 0$*, where* $\varepsilon_{\mathbf{i}}$ *has conditional density* $f(\cdot)$ *(possibly asymmetric or multimodal), and assume*

the regularity conditions C1–C7 hold. Define the kernel-based objective

$$Q_{\tilde{\mathbf{n}}}(\boldsymbol{\theta}) = \frac{1}{\tilde{\mathbf{n}} h_1 h_2^d} \sum_{\mathbf{i} \in \mathbf{I}_{\tilde{\mathbf{n}}}} \phi\left(\frac{Y_{\mathbf{i}} - \boldsymbol{\theta}^T \mathbf{X}_{\mathbf{i}}^*}{h_1}\right) K\left(\frac{\mathbf{X}_{\mathbf{i}} - \mathbf{x}}{h_2}\right),$$

with $\mathbf{X}_{\mathbf{i}}^* = [1 \ (\mathbf{X}_{\mathbf{i}} - \mathbf{x})^T / h_2]^T$. *Let* $\hat{\boldsymbol{\theta}}_{\text{MEM}}$ *and* $\hat{\boldsymbol{\theta}}_{\text{mean}}$ *denote the MEM and local linear mean estimators, respectively. Then, if* $m(\mathbf{x}) \neq \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$ *and* $f(\cdot)$ *is asymmetric or multimodal, we have*

$$Q_{\tilde{\mathbf{n}}}(\hat{\boldsymbol{\theta}}_{\text{MEM}}) \geq Q_{\tilde{\mathbf{n}}}(\hat{\boldsymbol{\theta}}_{\text{mean}}) + \delta_n,$$

with high probability, for some $\delta_n > 0$.

Proof. By definition, the MEM estimator satisfies $Q_{\tilde{\mathbf{n}}}(\hat{\boldsymbol{\theta}}_{\text{MEM}}) = \max_{\boldsymbol{\theta}} Q_{\tilde{\mathbf{n}}}(\boldsymbol{\theta})$, so we trivially have $Q_{\tilde{\mathbf{n}}}(\hat{\boldsymbol{\theta}}_{\text{MEM}}) \geq Q_{\tilde{\mathbf{n}}}(\hat{\boldsymbol{\theta}}_{\text{mean}})$. The inequality is strict unless $\hat{\boldsymbol{\theta}}_{\text{mean}}$ also maximizes $Q_{\tilde{\mathbf{n}}}(\cdot)$, which only occurs when the conditional distribution is symmetric and the mean coincides with the mode.

Now, assume that $m(\mathbf{x}) \neq \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$. Since $\phi(u)$ is strictly unimodal and symmetric around zero, it is maximized when the residual $Y_{\mathbf{i}} - \boldsymbol{\theta}^T \mathbf{X}_{\mathbf{i}}^*$ is close to zero. The MEM estimator $\hat{\boldsymbol{\theta}}_{\text{MEM}}$ targets the conditional mode and aligns residuals towards this central value more effectively than $\hat{\boldsymbol{\theta}}_{\text{mean}}$ when the mode and mean differ. For any observation $\mathbf{i}$ in the local window $\|\mathbf{X}_{\mathbf{i}} - \mathbf{x}\| \leq C h_2$, define the residuals $R_{\text{MEM},\mathbf{i}} := Y_{\mathbf{i}} - \hat{\boldsymbol{\theta}}_{\text{MEM}}^T \mathbf{X}_{\mathbf{i}}^*$ and $R_{\text{mean},\mathbf{i}} := Y_{\mathbf{i}} - \hat{\boldsymbol{\theta}}_{\text{mean}}^T \mathbf{X}_{\mathbf{i}}^*$. Since $\phi(u)$ is decreasing in $|u|$, and $|R_{\text{MEM},\mathbf{i}}| < |R_{\text{mean},\mathbf{i}}|$ for a nontrivial proportion of $\mathbf{i}$ due to mode–mean asymmetry, we have

$$\phi\left(\frac{R_{\mathrm{MEM},\mathbf{i}}}{h_1}\right) > \phi\left(\frac{R_{\mathrm{mean},\mathbf{i}}}{h_1}\right).$$

Let p_n denote the proportion of such $\mathbf{i}$ satisfying this inequality. Then, the aggregate gain in the objective satisfies

$$Q_{\tilde{\mathbf{n}}}(\hat{\boldsymbol{\theta}}_{\mathrm{MEM}}) - Q_{\tilde{\mathbf{n}}}(\hat{\boldsymbol{\theta}}_{\mathrm{mean}}) \geq \frac{p_n}{h_1 h_2^d}\left(\min_i K\left(\frac{\mathbf{X}_\mathbf{i} - \mathbf{x}}{h_2}\right)\right)\Delta_\phi,$$

where $\Delta_\phi = \min_{\mathbf{i}\in\mathcal{I}}\left[\phi\left(\frac{R_{\mathrm{MEM},\mathbf{i}}}{h_1}\right) - \phi\left(\frac{R_{\mathrm{mean},\mathbf{i}}}{h_1}\right)\right] > 0$. Under uniform consistency of both estimators and the regularity conditions of the kernel and bandwidths, we obtain $Q_{\tilde{\mathbf{n}}}(\hat{\boldsymbol{\theta}}_{\mathrm{MEM}}) > Q_{\tilde{\mathbf{n}}}(\hat{\boldsymbol{\theta}}_{\mathrm{mean}}) + \delta_n$, where $\delta_n = \Omega(p_n h_1^{-1} h_2^{-d})$ vanishes only if $m(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$, which contradicts the assumption. This completes the proof. □

Theorem S2 establishes that, under general non-Gaussian error structures, the MEM estimator yields a strictly higher value of the kernel-smoothed conditional likelihood than conventional local linear mean regression. This theoretical distinction is especially significant in scenarios characterized by asymmetry or multimodality in the response distribution, precisely the regimes in which mean-based estimators may provide a distorted or unrepresentative summary of the conditional behavior. The inequality in Theorem S2 underscores that, even in the absence of global optimality guarantees, the MEM estimator improves the fidelity of likelihood-based inference by targeting the dominant mode of the conditional distribution.

This lends strong theoretical support to the use of modal regression in practice. Furthermore, simulation results in Section 4 of the paper provide empirical validation, i.e., across a variety of sample sizes and distributional settings, the MEM estimator consistently attains higher values of the objective function than its mean-based counterpart, reinforcing its robustness and practical utility in mode-oriented analysis.

S2 Comments for Conditions in Subsection 2.3

The theoretical conditions listed in Subsection 2.3 are standard in the setting of local smoothers and modal regression models, necessary for establishing asymptotic properties, and their rationale can be elaborated upon. Condition C1 implicitly assumes that the mode of the density function $g(\cdot)$ is globally uniquely defined. In reality, enforcing the unique mode assumption is not imperative, suggesting that the technique described in this paper can also be applied to the multimode case to capture spatial clustering; see Ullah et al. (2022, 2023). Condition C2 specifies the requirements on the kernel function $K(\cdot)$ used, which are essential for technical reasons in the proofs and are common in nonparametric kernel estimation. The bounded support restriction on the kernel function is not indispensable and can be released by imposing a restriction on its tail. In particular, the standard multivari-

ate Gaussian kernel is allowed, i.e., $K(\mathbf{x}) = (2\pi)^{-d/2}\exp(-\mathbf{x}^T\mathbf{x}/2)$, which is infinitely supported and is the default kernel adopted in the numerical parts of this paper. Since $\phi(\cdot)$ is chosen as the Gaussian kernel, no specific conditions are listed for it. Condition C3 is utilized to regulate the variance term of the estimation, which has been employed, for instance, by Hallin et al. (2004) and Hallin et al. (2009). It suggests that the joint probability distribution $f_{\mathbf{i}_1,\cdots,\mathbf{i}_s}(\mathbf{X}_{\mathbf{i}_1},\cdots,\mathbf{X}_{\mathbf{i}_s})$ exists and is bounded uniformly in $\mathbf{i}_1,\cdots,\mathbf{i}_s$ for $s = 1,\cdots,2r-1$, in which $r \geq 1$ is a given integer. If the random field is composed of independent observations, then $|f_{\mathbf{i},\mathbf{j}}(\mathbf{x}',\mathbf{x}'') - f(\mathbf{x}')f(\mathbf{x}'')|$ vanishes as soon as $\mathbf{i}$ and $\mathbf{j}$ are distinct. Condition C3 governs local dependence through the distance between $f_{\mathbf{i},\mathbf{j}}(\mathbf{x}',\mathbf{x}'')$ and $f(\mathbf{x}')f(\mathbf{x}'')$, whereas the mixing condition regulates the dependence of sites which are far from each other through the distance between $P(AB)$ and $P(A)P(B)$. Note that the assumption of strict stationarity is a cornerstone in nonparametric regression estimation for spatial data and plays a fundamental role in this paper, ensuring that the conditional regression function depends solely on $\mathbf{X}$ and not on the specific site $\mathbf{i}$. This implies that for any nonnegative s in $\mathbb{Z}$ and any $\mathbf{j}^k = \left(j_1^k,\cdots,j_N^k\right)$ in $\mathbb{Z}^N$ with $k = 1,\cdots,s$, the joint probability measure of $\left(Y_{\mathbf{i}},X_{\mathbf{i}}\right),\left(Y_{\mathbf{i}+\mathbf{j}^1},X_{\mathbf{i}+\mathbf{j}^1}\right),\cdots,\left(Y_{\mathbf{i}+\mathbf{j}^s},X_{\mathbf{i}+\mathbf{j}^s}\right)$ remains the same regardless of the initial $\mathbf{i} = (i_1,\cdots,i_N)$ in $\mathbb{Z}^N$. While this assumption may not always hold

in environmental contexts, it can be relaxed by assuming that the $X_{\mathbf{i}}$'s are non-identically distributed, which is of particular interest to explore but is outside the scope of the present paper.

Condition C4 is regular in the literature of nonparametric regression to characterize the functional space of a model. The smoothness condition on the unknown function $m(\mathbf{x})$ determines the convergence rate of the spatial modal estimator. Higher-order smoothness is required if local polynomial estimation is applied. Condition C5 is employed for establishing asymptotic properties. Following the arguments presented in this paper, analogous results can be straightforwardly derived, in which $\varphi(\cdot)$ decays exponentially, i.e., $\varphi(t) \leq C \exp(-st)$ for some $s > 0$. Conditions C6-C7 are typical technical conditions in the spatial data context, which are essential for modal regression. They are crucial to achieving the same rate of convergence as in the independence case for the proposed spatial modal estimators. If we impose $\chi\left(n', n''\right) \leq C\left(n' + n'' + 1\right)^{\kappa}$ for some $C > 0$ and $\kappa > 1$ in condition C5, the last condition in C6 would be replaced by $(\tilde{\mathbf{n}}^{\kappa+1}/p)\varphi(q) \to 0$. In contrast to local linear spatial mean regression, we do not require the existence of moments for the error terms. However, to ensure the presence of asymptotic variance and bias of the estimator in the case of random fields with a spatial trend (see Remark 2.4 in the paper), we need to impose a condition

that the random variables $Y_{\mathbf{i}}$ and $\mathbf{X_i}$ possess finite absolute moments of order $(2+\delta)$, i.e., $\mathbb{E}[|Y_{\mathbf{i}}|^{2+\delta}] < \infty$ and $\mathbb{E}[\|\mathbf{X_i}\|^{2+\delta}] < \infty$ for some $\delta > 0$, which is the classic rank condition constraining covariate moments. The bandwidth significantly influences the behavior of the developed modal estimators, as is typical in all kernel estimations. Aside from condition C7, all other conditions related to bandwidths are delineated in the asymptotic theorems.

S3 Modal Boundary Behavior

The local linear approximation offers significantly improved boundary behavior compared to the local constant approach. It is natural to inquire whether the spatial modal estimators proposed in this paper maintain their asymptotic properties near the boundaries. For simplicity, we assume that there is a univariate regressor X $(d = 1)$ with a bounded support, i.e., $[-M, M]$. By employing an argument similar to the one developed in the proof of Theorem 2.2, it can be shown that asymptotic normality still holds near the boundary point $x = -M+ch_2$, where c is a positive constant. However, there are adjustments in the asymptotic biases and variances such that

$$B_0(x) = \frac{(\int_{-c}^{M} u^2 K(u)du)^2 - \int_{-c}^{M} uK(u)du \int_{-c}^{M} u^3 K(u)du}{\int_{-c}^{M} K(u)du \int_{-c}^{M} u^2 K(u)du - (\int_{-c}^{M} uK(u)du)^2} \left[\frac{h_2^2}{2} \frac{\partial^2 m(x)}{\partial x^2}\Big|_{x=-M^+}\right] - \frac{h_1^2}{2} \frac{(g^{(3)}(0 \mid -M^+))}{(g^{(2)}(0 \mid -M^+))},$$

$$B_1(x) = \frac{\int_{-c}^{M} K(u)du \int_{-c}^{M} u^3 K(u)du - \int_{-c}^{M} uK(u)du \int_{-c}^{M} u^2 K(u)du}{\int_{-c}^{M} K(u)du \int_{-c}^{M} u^2 K(u)du - (\int_{-c}^{M} uK(u)du)^2} \left[\frac{h_2^2}{2} \frac{\partial^2 m(x)}{\partial x^2}\Big|_{x=-M^+} \right],$$

$$\sigma_0^2(x) = \frac{\int \phi^2(t)t^2 dt}{n_l h_2 h_1^3 f(-M^+)} \frac{g(0 \mid -M^+)}{(g^{(2)}(0 \mid -M^+))^2} \frac{\mathcal{V}_0}{(\int_{-c}^{M} K(u)du \int_{-c}^{M} u^2 K(u)du - (\int_{-c}^{M} uK(u)du)^2)^2},$$

and

$$\sigma_1^2(x) = \frac{\int \phi^2(t)t^2 dt}{n_l h_2 h_1^3 f(-M^+)} \frac{g(0 \mid -M^+)}{(g^{(2)}(0 \mid -M^+))^2} \frac{\mathcal{V}_1}{(\int_{-c}^{M} K(u)du \int_{-c}^{M} u^2 K(u)du - (\int_{-c}^{M} uK(u)du)^2)^2},$$

respectively, where

$$\mathcal{V}_0 = \Big(\int_{-c}^{M} u^2 K(u)du \Big)^2 \int_{-c}^{M} K^2(u)du - 2 \int_{-c}^{M} uK(u)du \int_{-c}^{M} u^2 K(u)du \int_{-c}^{M} uK^2(u)du + \Big(\int_{-c}^{M} uK(u)du \Big)^2 \int_{-c}^{M} u^2 K^2(u)du,$$

and

$$\mathcal{V}_1 = \Big(\int_{-c}^{M} uK(u)du \Big)^2 \int_{-c}^{M} K^2(u)du - 2 \int_{-c}^{M} uK(u)du \int_{-c}^{M} K(u)du \int_{-c}^{M} uK^2(u)du + \Big(\int_{-c}^{M} K(u)du \Big)^2 \int_{-c}^{M} u^2 K^2(u)du.$$

Indeed, this boundary advantage would likely become more pronounced as N grows. Consequently, local linear modal estimation exhibits automatic good behavior at boundaries without the need for boundary correction. This holds true for both the left boundary point $x = -M + ch_2$ and the right

boundary point $x = M - ch_2$. Even if point M were an interior point, the same results would still apply with $c = M$.

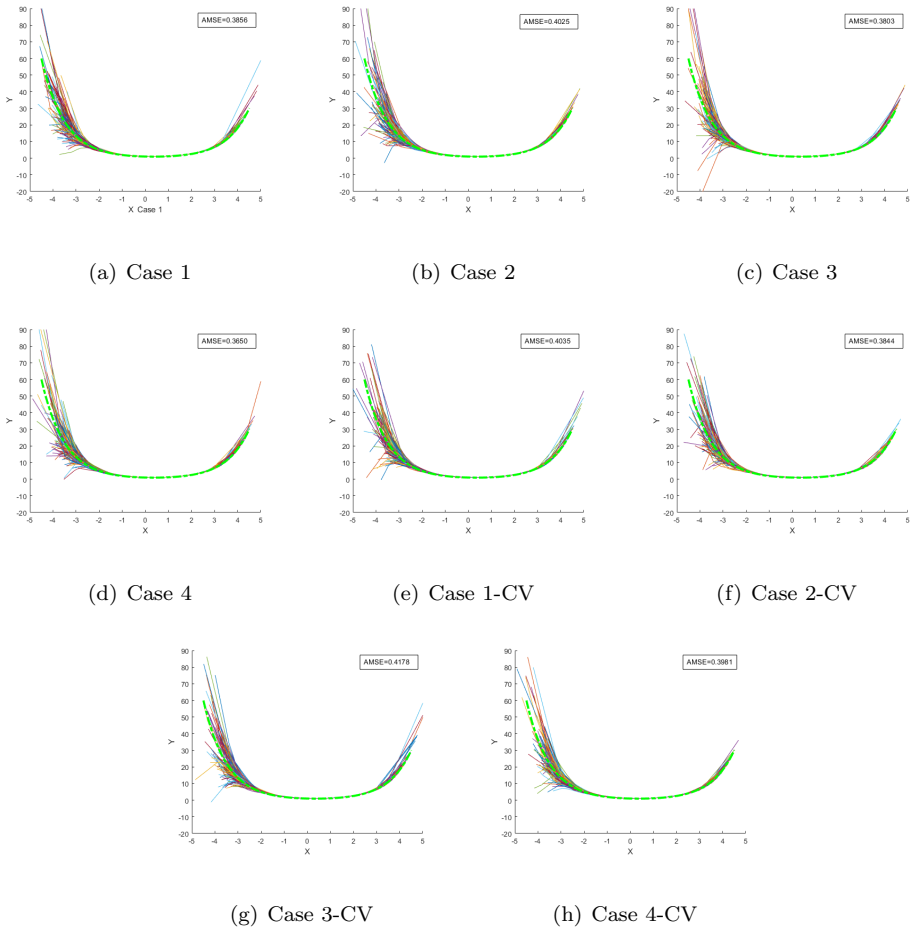
S4 Additional Monte Carlo Experiments

DGP 1 (Asymmetric Data) For DGP 1 listed in the paper, we also investigate the impact of including additional spatial lags into the definition of $X_{i,j}$ on the estimation of spatial modal regression. For illustration, we specifically concentrate on the scenario when $(n_1, n_2) = (20, 30)$, while the results for other sample sizes are comparable. We adopt the same model as in DGP 1 but generate $X_{i,j}$ according to the following four different equations

$$\begin{cases} \text{Case 1: } X_{i,j} = \sin(X_{i-1,j} + X_{i,j-1}) + e_{i,j}, \\ \text{Case 2: } X_{i,j} = \sin(X_{i+1,j} + X_{i,j+1}) + e_{i,j}, \\ \text{Case 3: } X_{i,j} = \sin(X_{i-2,j} + X_{i,j-2} + X_{i-1,j} + X_{i,j-1}) + e_{i,j}, \\ \text{Case 4: } X_{i,j} = \sin(X_{i-2,j} + X_{i,j-2} + X_{i-1,j} + X_{i,j-1} + X_{i+1,j} \\ \qquad\qquad +X_{i,j+1} + X_{i+2,j} + X_{i,j+2}) + e_{i,j}. \end{cases}$$

Compared to Figure 1-(d), the results in Figure S1 indicate that the estimated spatial modal regression is not sensitive to the choice of lags, implying that the lags of $X_{i,j}$ have little influence on estimation. These findings are consistent with those obtained by Hallin et al. (2004) in spatial mean regression. We also report the simulation results in which the band-

width is obtained from the modal cross-validation (CV) procedure. The results demonstrate that the suggested data-based "rule of thumb" bandwidth choice procedure performs well, as evidenced by nearly identical AMSEs.



(a) Case 1 (b) Case 2 (c) Case 3

(d) Case 4 (e) Case 1-CV (f) Case 2-CV

(g) Case 3-CV (h) Case 4-CV

Figure S1: Impact of Spatial Lags on Modal Estimator of Function $m(\cdot)$

DGP 2 (Symmetric Data) To illustrate the robustness performance of the developed spatial modal estimator, we generate data from the model

$$Y_{i,j} = m(X_{i,j}) + \varepsilon_{i,j} \text{ with } m(x) = \sin(\pi x),$$

where $\{\varepsilon_{i,j}, (i,j) \in \mathbb{Z}^2\}$ are independently taken from one of the following distributions: (i) standard normal distribution $\mathcal{N}(0,1)$; (ii) fat tails t distribution with degrees of freedom equal to 3, i.e., $t(3)$; (iii) symmetric heavy-tailed mixture Laplace distribution $0.8L_p(0,1)+0.2L_p(0,5)$; (iv) symmetric contaminated normal distribution $0.9\mathcal{N}(0,1)+0.1\mathcal{N}(0,10^2)$, and $\{X_{i,j}, (i,j) \in \mathbb{Z}^2\}$ are produced by the spatial autoregression $X_{i,j} = \sin(X_{i-1,j} + X_{i,j-1} + X_{i+1,j} + X_{i,j+1}) + e_{i,j}$ with $\{e_{i,j}, (i,j) \in \mathbb{Z}^2\} \sim \mathcal{N}(0,1)$. We then have $Mode(Y_{i,j} \mid X_{i,j}) = \mathbb{E}(Y_{i,j} \mid X_{i,j}) = \sin(\pi X_{i,j})$. The data simulation setup mirrors that of DGP 1, ensuring the generation of stationary points. We also conduct simulations across four different sample sizes, i.e., $(n_1, n_2) = (10,10), (15,15)$, (20, 20), and (20,30), separately, and provide the associated AMSEs based on 200 replications. The bandwidth for spatial mean regression is chosen by the traditional cross-validation method built on the MSE criterion.

Table S1: Results of Simulations—DGP 2

Distribution	(n_1, n_2)	Modal	Local Linear	(n_1, n_2)	Modal	Local Linear
$\mathcal{N}(0,1)$	(10,10)	0.1445	0.1243	(15,15)	0.1087	0.0866
	(20,20)	0.0845	0.0743	(20,30)	0.0713	0.0666
$t(3)$	(10,10)	0.2091	0.2450	(15,15)	0.1587	0.1854
	(20,20)	0.1247	0.1353	(20,30)	0.1027	0.1173
$0.8L_p(0,1) + 0.2L_p(0,5)$	(10,10)	0.3197	0.3369	(15,15)	0.2239	0.2689
	(20,20)	0.1867	0.2347	(20,30)	0.1750	0.2184
$0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(0,10^2)$	(10,10)	0.2570	0.2814	(15,15)	0.1830	0.2314
	(20,20)	0.1639	0.2144	(20,30)	0.1517	0.2071

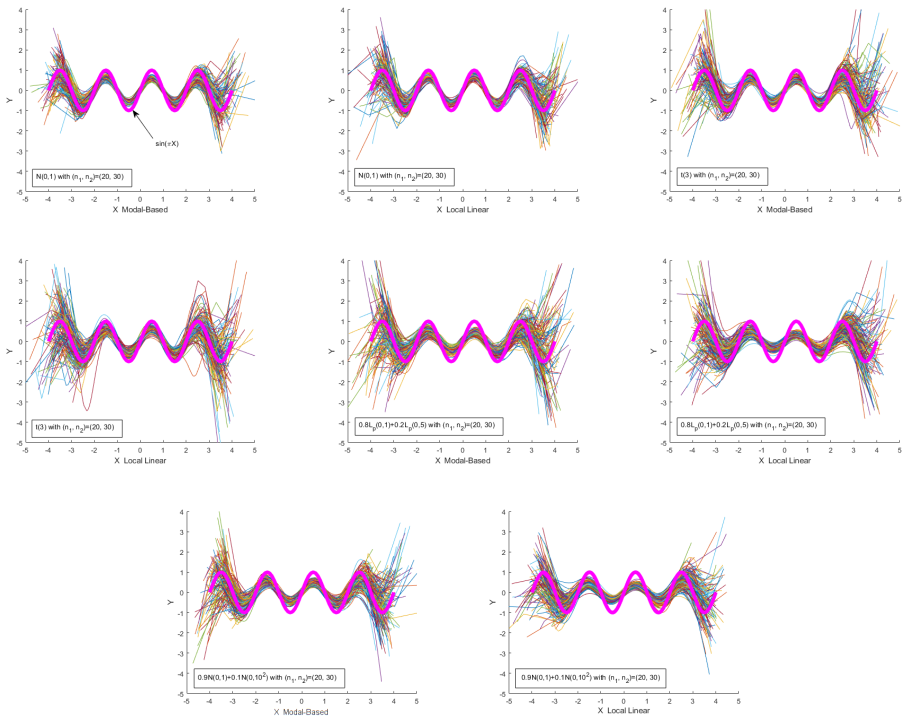


Figure S2: Modal and Mean Estimators of Function $m(\cdot)$

The simulation results in terms of AMSE are presented in Table S1 and Figure S2. Because of the similarity of observations, only results for $(n_1, n_2) = (20, 30)$ are reported in the figure. In Figure S2, the pink solid line denotes the true spatial regression function. The results show that the proposed modal estimation procedure is on average as good as or better than local linear mean estimation method (represented by "Local Linear"). From Table S1, we can see that both modal and mean estimations exhibit satisfactory performance for errors following a standard normal distribution. Also, local linear mean estimation slightly outperforms modal estimation

somewhat in such a case. However, in the presence of contaminated or heavy-tailed errors, modal estimation significantly outperforms local linear mean estimation across all sample instances. This trend is consistent with the observations in Figure S2. These results underscore the robustness and efficiency of the proposed spatial modal estimator compared to the spatial mean estimator. Furthermore, in line with the asymptotic theory outlined in Section 2, all AMSEs decrease with increasing sample size.

S5 Additive Spatial Modal Regression

As elucidated in Section 1, the implementation of the proposed nonparametric spatial modal regression becomes progressively intricate with an increasing number of covariates. An alternative strategy is to employ an additive modal regression model, designed to mitigate the curse of dimensionality and uncover crucial features often overlooked by additive mean or quantile regression models. The additive models, on the other hand, are well-suited for approximating the conditional mode of a spatial random variable given its nearest neighbor observations. Despite the extensive literature and recent advancements in additive models, as far as we know, the statistical estimation issue in additive spatial modal regression remains unexplored. Leveraging the developed estimation method from the previous sections, we

delve into the exploration of how to estimate the optimal additive approximation for the conditional spatial modal function (1.1). Specifically, the additive model decomposes $m(\mathbf{X})$ into an additive sum of the form

$$m(\mathbf{X}) = \mu + \sum_{\kappa=1}^{d} m_\kappa(X_\kappa), \tag{S1.1}$$

where μ represents the spatial modal regression intercept, defined as $Mode(Y - \sum_{\kappa=1}^{d} m_\kappa(X_\kappa)) = \mu$, and the functions $m_1(\cdot), \cdots, m_d(\cdot)$ are real measurable functions valued in the space functions $\mathcal{M} = \{m_\kappa \in C^2(\mathbb{R}) : \|m_\kappa\| = \sup_{X_\kappa \in \mathbb{R}} |m_\kappa(X_\kappa)| < C\}$, with $C^2(\mathbb{R})$ the space of twice differentiable functions and C a positive constant. For identification, location normalization is required. We assume that all mode values of $m_\kappa(X_\kappa)$ are zero, that is, $Mode(m_\kappa(X_\kappa)) = 0$, without imposing any other conditions on moments. Otherwise, we set

$$\tilde{m}_\kappa(X_\kappa) = m_\kappa(X_\kappa) - Mode(m_\kappa(X_\kappa)) \text{ and } \tilde{\mu} = \mu + \sum_{\kappa=1}^{d} Mode(m_\kappa(X_\kappa)). \tag{S1.2}$$

In such a scenario, the optimal convergence rate shall remain the same as that in the univariate nonparametric spatial modal regression under certain regularity conditions, regardless of the value of d. To numerically solve the

The mode identification condition for the developed additive spatial modal regression model is different from that of the additive spatial mean and quantile regression models, namely, $\mathbb{E}(m_\kappa(X_\kappa)) = 0$ and $Q_\tau(m_\kappa(X_\kappa)) = 0$, where $Q_\tau(\cdot)$ represents the τth quantile ($\tau \in [0, 1]$). However, if the moments exist for the suggested additive spatial modal regression, we suppose that the mean identification condition could also be utilized for identifying the proposed model.

developed additive spatial modal regression, we can suggest a nonparametric kernel-based backfitting algorithm.

Based on Definition 2.1, the optimal additive approximation is derived by maximizing

$$\mathbb{E}\left[\frac{1}{h_3}\phi\left(\frac{Y_{\mathbf{i}}-\mu-\sum_{\kappa=1}^{d}m_\kappa(X_{\mathbf{i}\kappa})}{h_3}\right)\right] \tag{S1.3}$$

over $\mu+\sum_{\kappa=1}^{d}m_\kappa(X_{\mathbf{i}\kappa})\in\mathcal{F}_{add}$, where $\mathcal{F}_{add}=\{\mu+\sum_{\kappa=1}^{d}m_\kappa(X_{\mathbf{i}\kappa})\mid \mu\in\mathbb{R}, Mode(m_\kappa(X_{\mathbf{i}\kappa}))=0 \text{ for } 1\leq\kappa\leq d\}$. However, directly extending the local linear estimator to the additive spatial modal regression is challenging due to the presence of nuisance functions $m_\iota(X_{\mathbf{i}\iota})$'s for $\iota\neq\kappa$ when estimating $m_\kappa(X_{\mathbf{i}\kappa})$. To address this, we extend the findings of Yu and Lu (2004), who developed an estimator for the components of a nonparametric additive quantile regression, to treat other components of the model as known when one of them is estimated.

To be more specific, we define the fitted value of (S1.1) at the point $\mathbf{x}$ as

$$\hat{m}(\mathbf{x})=\hat{\mu}+\sum_{\kappa=1}^{d}\hat{m}_\kappa(x_\kappa), \tag{S1.4}$$

in which

$$\hat{\mu}=\arg\max_{\mu}\frac{1}{\tilde{\mathbf{n}}h_3}\sum_{\mathbf{i}\in I_{\mathbf{n}}}\phi\left(\frac{Y_{\mathbf{i}}-\mu-\sum_{\kappa=1}^{d}m_\kappa(X_{\mathbf{i}\kappa})}{h_3}\right) \tag{S1.5}$$

with the precondition that all component functions $\{m_\kappa(X_{\mathbf{i}\kappa})\}_{k=1}^{d}$ are known, where $h_3=h_3(\mathbf{n})$ is a bandwidth that depends on $\mathbf{n}$ and approaches zero as

$\mathbf{n} \to \infty$, $\hat{m}_\kappa(x_\kappa) = \hat{a}_\kappa$, and $\hat{m}_\kappa^{(1)}(x_\kappa) = \hat{b}_\kappa$. The values $\hat{a}_\kappa$ and $\hat{b}_\kappa$ are obtained by maximizing the following kernel-based objective function under the precondition that μ and other component functions $\{m_\iota(X_{\mathbf{i}\iota})\}_{\iota\neq\kappa,\iota=1}^{d}$ are known

$$Q(a_\kappa, b_\kappa) = \frac{1}{\tilde{\mathbf{n}} h_4 h_{5\kappa}} \sum_{\mathbf{i}\in I_\mathbf{n}} \phi\left(\frac{Y_\mathbf{i} - \mu - \sum_{\iota\neq\kappa,\iota=1}^{d} m_\iota(X_{\mathbf{i}\iota}) - a_\kappa - b_\kappa(X_{\mathbf{i}\kappa} - x_\kappa)}{h_4}\right) K\left(\frac{X_{\mathbf{i}\kappa} - x_\kappa}{h_{5\kappa}}\right) \tag{S1.6}$$

using a local linear approximation for $m_\kappa(X_{\mathbf{i}\kappa})$, i.e., $m_\kappa(X_{\mathbf{i}\kappa}) \approx m_\kappa(x_\kappa) + m_\kappa^{(1)}(x_\kappa)(X_{\mathbf{i}\kappa} - x_\kappa) = a_\kappa + b_\kappa(X_{\mathbf{i}\kappa} - x_\kappa)$, and h_4 and $h_{5\kappa}$ are two bandwidths that depend on $\mathbf{n}$ and approach zero as $\mathbf{n} \to \infty$. For flexibility, we let $h_{5\kappa}$ be different for each independent variable $X_{\mathbf{i}\kappa}$. As discussed later, the cautious choice of bandwidths is crucial for the proposed estimators to be asymptotically normal. The practical selection of bandwidths for the suggested additive spatial modal regression shall be investigated in the end.

In reality, the oracle equations (S1.5) and (S1.6) cannot be directly utilized for estimation due to the lack of knowledge regarding the true values of the other components when estimating $m_\kappa(X_{\mathbf{i}\kappa})$. To overcome this challenge, we employ a backfitting algorithm in conjunction with the MEM algorithm. Specifically, for the numerical estimation of the proposed additive spatial modal regression, we devise a kernel-based backfitting algorithm by preestimating components used to construct oracle estimators as if they

Algorithm 0: Backfitting-MEM Algorithm

Step-1-Initial Estimation. Estimate $\hat{\mu}^{(0)} = \arg\max_{\mu} \frac{1}{\tilde{\mathbf{n}} h_3} \sum_{\mathbf{i} \in I_{\mathbf{n}}} \phi\left(\frac{Y_{\mathbf{i}} - \mu}{h_3}\right)$ and

$$(\hat{a}_\kappa, \hat{b}_\kappa) = \arg\max_{a_\kappa, b_\kappa} \frac{1}{\tilde{\mathbf{n}} h_4 h_{5\kappa}} \sum_{\mathbf{i} \in I_{\mathbf{n}}} \phi\left(\frac{Y_{\mathbf{i}} - \hat{\mu}^{(0)} - a_\kappa - b_\kappa(X_{\mathbf{i}\kappa} - x_\kappa)}{h_4}\right) K\left(\frac{X_{\mathbf{i}\kappa} - x_\kappa}{h_{5\kappa}}\right)$$

for $\kappa = 1, \cdots, d$. Then, set $m_\kappa^{(0)}(x_\kappa) = \hat{a}_\kappa$ and take $m_\kappa^{*(0)}(x_\kappa)$ as $m_\kappa^{(0)}(x_\kappa) - Mode(\{m_\kappa^{(0)}(X_{\mathbf{i}\kappa})\}_{\mathbf{i} \in I_{\mathbf{n}}})$.

Step-2-Iteration. Set $\hat{\mu}^{(g)} = \arg\max_{\mu} \frac{1}{\tilde{\mathbf{n}} h_4} \sum_{\mathbf{i} \in I_{\mathbf{n}}} \phi\left(\frac{Y_{\mathbf{i}} - \sum_{\kappa=1}^{d} m_\kappa^{*(g-1)}(X_{\mathbf{i}\kappa}) - \mu}{h_4}\right)$, where g is the iteration indicator, and

$$(\hat{a}_\kappa, \hat{b}_\kappa) = \arg\max_{a_\kappa, b_\kappa} \frac{1}{\tilde{\mathbf{n}} h_4 h_{5\kappa}} \sum_{\mathbf{i} \in I_{\mathbf{n}}} \phi\left(\frac{Y_{\mathbf{i}} - \hat{\mu}^{(g)} - \sum_{\iota \neq \kappa, \iota=1}^{d} m_\iota^{*(g-1)}(X_{\mathbf{i}\iota}) - a_\kappa - b_\kappa(X_{\mathbf{i}\kappa} - x_\kappa)}{h_4}\right) K\left(\frac{X_{\mathbf{i}\kappa} - x_\kappa}{h_{5\kappa}}\right).$$

Let $m_\kappa^{(g)}(x_\kappa) = \hat{a}_\kappa$ and take $m_\kappa^{*(g)}(x_\kappa)$ as $m_\kappa^{(g)}(x_\kappa) - Mode(\{m_\kappa^{(g)}(X_{\mathbf{i}\kappa})\}_{\mathbf{i} \in I_{\mathbf{n}}})$.

Iterate-Updating. Keep iterating Step-2 for $g = 1, 2, 3, \cdots$ until the value of $(\hat{\mu}^{(g)}, m_1^{*(g)}, m_2^{*(g)}, \cdots, m_d^{*(g)})$ has converged. Next, for $\kappa = 1, \cdots, d$, set $(\hat{a}_\kappa, \hat{b}_\kappa) = (m_\kappa^{*(g)}(x_\kappa), \hat{b}_\kappa)$. The final estimators $(\hat{a}_\kappa, \hat{b}_\kappa) = (\hat{m}_\kappa(x_\kappa), \hat{m}_\kappa^{(1)}(x_\kappa))$.

were true functions. It can be seen from Algorithm 0 that preliminary estimates are updated after each additive component is estimated. Meanwhile, in each step, only one additive component is revised through the MEM algorithm, leaving the other components fixed. As a result, the proposed estimation procedure is computationally expedient. According to the mapping theorem, the developed backfitting-MEM algorithm is a first-order algorithm with similar properties as the classical backfitting algorithm, which

may result in sluggard convergence. However, by utilizing the Gaussian kernel for $\phi(\cdot)$ during modal estimation, we obtain an explicit expression in the M-Step, leading to reduced computational costs and faster convergence.

As discussed earlier, a significant benefit of the developed additive spatial modal regression model lies in its capability to achieve a one-dimensional convergence rate. For simplicity, we primarily highlight the convergence rate and asymptotic normality of the oracle estimators, along with offering some insights into the asymptotic results derived from Algorithm 0.

Theorem S1.1. *Under the regularity conditions C1-C7 (with $d = 1$ and $X_{\mathbf{i}\kappa}$ and $K(\cdot)$ be univariate), with probability approaching one, as $\tilde{\mathbf{n}} \to \infty$, $h_4 \to 0$, $h_{5\kappa} \to 0$, and $\tilde{\mathbf{n}} h_{5\kappa} h_4^5 \to \infty$, there exist consistent maximizers $(\hat{m}_\kappa(x_\kappa), \hat{m}_\kappa^{(1)}(x_\kappa))$ of* (S1.6) *such that*

i. $|\hat{m}_\kappa(x_\kappa) - m_\kappa(x_\kappa)| = O_p\left((\tilde{\mathbf{n}} h_{5\kappa} h_4^3)^{-1/2} + h_4^2 + h_{5\kappa}^2\right)$,

ii. $|h_{5\kappa}(\hat{m}_\kappa^{(1)}(x_\kappa) - m_\kappa^{(1)}(x_\kappa))| = O_p\left((\tilde{\mathbf{n}} h_{5\kappa} h_4^3)^{-1/2} + h_4^2 + h_{5\kappa}^2\right)$,

where $m_\kappa(x_\kappa)$ is the spatial modal function of $Y - \mu - \sum_{\iota \neq \kappa, \iota = 1} m_\iota(X_{\mathbf{i}\iota})$ given $X_{\mathbf{i}\kappa} = x_\kappa$.

Theorem S1.2. *With $\tilde{\mathbf{n}} h_{5\kappa}^5 h_4^3 = O(1)$ and $\tilde{\mathbf{n}} h_{5\kappa} h_4^7 = O(1)$, under the same conditions as Theorem S1.1, if $n_k \left(h_5 h_{5\kappa}\right)^{\delta/(2+\delta)a} > 1$ for all $k = 1, \cdots, N$ as $\mathbf{n} \to \infty$, the estimators satisfying the consistency results in Theorem*

S1.1 have the following asymptotic result

$$\sqrt{\tilde{\mathbf{n}}h_{5\kappa}h_4^3}\Bigg[\begin{pmatrix}\hat{m}_\kappa(x_\kappa)-m_\kappa(x_\kappa)\\ h_{5\kappa}(\hat{m}_\kappa^{(1)}(x_\kappa)-m_\kappa^{(1)}(x_\kappa))\end{pmatrix}$$

$$-S^{*-1}\left(\frac{h_{5\kappa}^2}{2}m_\kappa^{(2)}(x_\kappa)\Lambda^*-\frac{h_4^2}{2}\frac{g^{(3)}(0\mid X_{\mathbf{i}\kappa}=x_\kappa)}{g^{(2)}(0\mid X_{\mathbf{i}\kappa}=x_\kappa)}\Gamma^*\right)\Bigg]$$

$$\xrightarrow{d}\mathcal{N}\left(0,\int\phi^2(t)t^2dt\frac{g(0\mid X_{\mathbf{i}\kappa}=x_\kappa)}{(g^{(2)}(0\mid X_{\mathbf{i}\kappa}=x_\kappa))^2}(f(x_\kappa))^{-1}S^{*-1}\Sigma^*S^{*-1}\right).$$

If we allow $\tilde{\mathbf{n}}h_{5\kappa}^5h_4^3\to 0$ *and* $\tilde{\mathbf{n}}h_{5\kappa}h_4^7\to 0$*, the asymptotic theorem becomes*

$$\sqrt{\tilde{\mathbf{n}}h_{5\kappa}h_4^3}\begin{pmatrix}\hat{m}_\kappa(x_\kappa)-m_\kappa(x_\kappa)\\ h_{5\kappa}(\hat{m}_\kappa^{(1)}(x_\kappa)-m_\kappa^{(1)}(x_\kappa))\end{pmatrix}$$

$$\xrightarrow{d}\mathcal{N}\left(0,\int\phi^2(t)t^2dt\frac{g(0\mid X_{\mathbf{i}\kappa}=x_\kappa)}{(g^{(2)}(0\mid X_{\mathbf{i}\kappa}=x_\kappa))^2}(f(x_\kappa))^{-1}S^{*-1}\Sigma^*S^{*-1}\right),$$

where $S^*=\begin{pmatrix}\int K(u)du & \int uK(u)du\\ \int uK(u)du & \int u^2K(u)du\end{pmatrix}$, $\Gamma^*=\begin{pmatrix}\int K(u)du\\ \int uK(u)du\end{pmatrix}$,

$\Lambda^*=\begin{pmatrix}\int u^2K(u)du\\ \int u^3K(u)du\end{pmatrix}$, *and* $\Sigma^*=\begin{pmatrix}\int K^2(u)du & \int uK^2(u)du\\ \int uK^2(u)du & \int u^2K^2(u)du\end{pmatrix}$.

Because additive models draw inspiration from local linear spatial modal regression, it is unsurprising that the proofs of the preceding two theorems closely follow the lines of Theorems 2.1 and 2.2. Therefore, we omit these proofs in this paper. Theorem S1.2 illustrates that the proposed method is adept at estimating the additive components at a one-dimensional non-

parametric optimal rate, irrespective of the size of d. Consequently, neither the dimension d nor the other function components influence the formation of the bandwidths for $m_\kappa(x\kappa)$. If a symmetric condition is imposed on the kernel function $K(\cdot)$, the estimators $\hat{m}_\kappa(x_\kappa)$ and $\hat{m}_\kappa^{(1)}(x_\kappa)$ will be asymptotically independent. Notice that we employ local linear approximation for all component functions. However, if various components are known to have different orders of differentiation, the rate of bias for all estimated functions will be determined by the rate of the lowest degree local polynomial. In such cases, if the primary interest is estimating the overall model, the backfitting estimator may converge more slowly than the one-dimensional nonparametric estimator. Although Theorem S1.2 is established only for the interior point, it is expected that the suggested estimator will not require boundary adjustments. Finally, we emphasize that the estimator described here can be easily extended to a generalized spatial additive modal regression model with a specified link function.

It is worth mentioning that the appearance of $\hat{\mu}$ in the objective function in Algorithm 0-Step-1 does not affect the asymptotic results for $(\hat{m}_\kappa(x_\kappa)$, $\hat{m}_\kappa^{(1)}(x_\kappa))$ owing to faster convergence rate. We can easily demonstrate that

$$|\hat{\mu} - \mu_0| = O_p\left((\tilde{\mathbf{n}}h_3^3)^{-1/2} + h_3^2\right), \tag{S1.7}$$

where μ_0 is the true value of μ. As a result, given the optimal bandwidths

obtained by minimizing AsyMSE, the above statement is automatically fulfilled. We can also show that the proposed estimator in Algorithm 0-Step-2 is asymptotically equivalent to $m_\kappa(x_\kappa)$. Thus, when estimating an additive component such as $m_\kappa(x_\kappa)$ in the additive modal regression model, we can achieve asymptotic performance as if the other additive components were known by undersmoothing. The fundamental idea is to choose a relative smaller bandwidth, ensuring that the bias of estimating other functions is asymptotically negligible when estimating $m_\kappa(x_\kappa)$. Since the asymptotic variance of the estimator is independent of the bandwidths associated with other functions, the bandwidths should be chosen to be as small as possible when estimating $m_\kappa(x_\kappa)$, which is the benefit of utilizing the kernel-based backfitting algorithm–avoiding data-based bandwidth selection for $m_\kappa(x_\kappa)$.

To select appropriate bandwidths for the proposed additive spatial modal regression model in empirical settings, several strategies are available. One practical approach is to apply the plug-in rule outlined in Subsection 2.4, which derives optimal bandwidths by minimizing the asymptotic mean squared error. Alternatively, one may adopt a data-driven selection method using a multi-stage cross-validation scheme, such as the concatenated cross-validation procedure proposed by Feng et al. (2020). Given our focus on estimating conditional modes rather than conditional means or

quantiles, conventional bandwidth selection criteria based on mean squared error or absolute deviation loss may not be appropriate. Instead, we propose a kernel-based cross-validation criterion tailored to modal estimation. Specifically, consider a validation set $\{(\mathbf{X_i}, Y_\mathbf{i})\}_{\mathbf{i}\in I_{\tilde{\mathbf{m}}}}$, and define the validation loss using a Gaussian kernel centered at the predicted mode $\hat{\mu}_\mathbf{i}$ for each observation. The goal is to select the bandwidth h_3 that maximizes the average modal fit, measured by

$$\arg\max_{h_3} \frac{1}{\tilde{\mathbf{m}} h_3} \sum_{\mathbf{i}\in I_{\tilde{\mathbf{m}}}} \exp\left(-\frac{(Y_\mathbf{i} - \hat{\mu}_\mathbf{i})^2}{h_3^2}\right). \tag{S1.8}$$

To implement this in practice, we begin by setting the initial value of h_3 to $\tilde{\mathbf{m}}^{-1/7}$, which corresponds to the AsyMSE-optimal bandwidth rate for local linear modal regression. We then conduct a first round of five-fold cross-validation to update and refine the estimate of h_3, yielding an intermediate value denoted by h_3^*. We repeat the procedure described above using a second five-fold cross-validation method to achieve the optimal value of h_3

$$\arg\max_{h_3} \frac{1}{\tilde{\mathbf{m}} h_3^*} \sum_{\mathbf{i}\in I_{\tilde{\mathbf{m}}}} \exp\left(-\frac{(Y_\mathbf{i} - \hat{\mu}_\mathbf{i})^2}{h_3^{*2}}\right). \tag{S1.9}$$

The best one is denoted as h_3^{**}. We finally implement a third five-fold cross-validation procedure

$$\arg\max_{h_3} \frac{1}{\tilde{\mathbf{m}} h_3^{**}} \sum_{\mathbf{i}\in I_{\tilde{\mathbf{m}}}} \exp\left(-\frac{(Y_\mathbf{i} - \hat{\mu}_\mathbf{i})^2}{h_3^{**2}}\right) \tag{S1.10}$$

to achieve the optimal value for h_3. This three-stage refinement procedure is designed to stabilize the selection process and mitigate sensitivity to local fluctuations in the validation loss surface. A similar procedure can be applied to select h_4 and $h_{5\kappa}$ by setting $h_4 = h_{5\kappa}$ for convenience. The initial values for h_4 and $h_{5\kappa}$ are both set to be $\tilde{\mathbf{m}}^{-1/8}$.

S6 Extended Models

The methodology developed in this paper encompasses a wide range of spatial modal regression models, including both parametric and nonparametric approaches. Several of them are listed below.

S6-1 Parametric Linear Spatial Modal Regression

As mentioned in the Introduction section, expanding the number of variables in nonparametric spatial modal regression becomes impractical due to the curse of dimensionality. When there is prior information to reasonably impose a parametric specification on the function $m(\cdot)$, such as $m(\mathbf{X}) = a + \mathbf{b}^T\mathbf{X}$, the kernel-based objective function becomes

$$Q_{\tilde{\mathbf{n}}}(a, \mathbf{b}) = \frac{1}{\tilde{\mathbf{n}} h_1} \sum_{\mathbf{i} \in I_{\mathbf{n}}} \phi\left(\frac{Y_{\mathbf{i}} - a - \mathbf{b}^T\mathbf{X}_{\mathbf{i}}}{h_1}\right),$$

referred to as the *linear spatial modal regression*. Following similar arguments as in the paper, it can be demonstrated that the linear spatial modal

estimators, $\hat{\boldsymbol{\theta}}_{LM} = (\hat{a}_{LM}, \hat{\mathbf{b}}_{LM}^T)^T$, are asymptotically distributed with a convergence rate of $\sqrt{\tilde{\mathbf{n}}h_1^3}$ under certain appropriate conditions

$$\sqrt{\tilde{\mathbf{n}}h_1^3}\left[\hat{\boldsymbol{\theta}}_{LM} - \boldsymbol{\theta}_{LM,0} - \frac{h_1^2}{2}(\mathbb{E}(g^{(2)}(0 \mid \mathbf{X_i}))\mathbf{X_i}\mathbf{X_i}^T)^{-1}\left(\mathbb{E}(g^{(3)}(0 \mid \mathbf{X_i}))\mathbf{X_i}\right)\right.$$
$$\left. \xrightarrow{d} \mathcal{N}\left(0, \int \phi^2(t)t^2dt\Sigma_\theta\right),\right.$$

where $\boldsymbol{\theta}_{LM,0}$ denotes the true parameter vector and

$$\Sigma_\theta = \left(\mathbb{E}(g^{(2)}(0 \mid \mathbf{X_i}))\mathbf{X_i}\mathbf{X_i}^T\right)^{-1}(\mathbb{E}(g(0 \mid \mathbf{X_i}))\mathbf{X_i}\mathbf{X_i}^T)^{-1}\left(\mathbb{E}(g^{(2)}(0 \mid \mathbf{X_i}))\mathbf{X_i}\mathbf{X_i}^T\right)^{-1}.$$

S6-2 Varying Coefficient Spatial Modal Regression

If we are concerned with estimating spatial modal regression with functional coefficients that smoothly vary over another covariate, we can define a varying coefficient spatial modal regression as $Y_\mathbf{i} = X_{\mathbf{i}1}\beta_1(\mathbf{U_i}) + \cdots + X_{\mathbf{i}d}\beta_d(\mathbf{U_i}) + \varepsilon_\mathbf{i}$ with $Mode(\varepsilon_\mathbf{i} \mid \mathbf{X_i}, \mathbf{U_i}) = 0$, where $Y_\mathbf{i}$, with values in R, $\mathbf{X_i}$, with values in R^d, and $\mathbf{U_i}$, with values in R^κ, are defined over a probability space $(\Omega, \mathcal{F}, P)$. Utilizing the concept of local linear fitting and assuming that each of the coefficients $\beta_r(\mathbf{U_i})$, $r = 1, \cdots, d$, is at least second continuously differentiable, we can solve the following maximization problem to obtain the corresponding spatial modal estimates

$$\max_{a_r, \mathbf{b}_r} \frac{1}{\tilde{\mathbf{n}}h_1h_2^k}\sum_{\mathbf{i}\in I_\mathbf{n}}\phi\left(\frac{Y_\mathbf{i} - \sum_{r=1}^d\left[\alpha_r + (\mathbf{U_i} - \mathbf{u}_0)^T\boldsymbol{b}_r\right]X_{\mathbf{i}r}}{h_1}\right)K\left(\frac{\mathbf{U_i} - \mathbf{u}_0}{h_2}\right),$$

where $\beta_r(\mathbf{U_i}) \approx \alpha_r + (\mathbf{U_i} - \mathbf{u})^T\boldsymbol{b}_r$ for $\mathbf{U_i}$ in the neighborhood of $\mathbf{u}$.

To mitigate the curse of dimensionality, it is common to set $\kappa = 1$. Let the corresponding spatial modal estimators be denoted as $(\tilde{\boldsymbol{\alpha}}(u), \tilde{\boldsymbol{b}}(u))$ with $\tilde{\boldsymbol{\alpha}}(u) = (\tilde{\alpha}_1(u), \cdots, \tilde{\alpha}_d(u))^T$ and $\tilde{\boldsymbol{b}}(u) = (\tilde{b}_1(u), \cdots, \tilde{b}_d(u))^T$. The true estimators are denoted as $(\boldsymbol{\alpha}_0(u), \boldsymbol{b}_0(u))$, $f_U(\cdot)$ is the density of U, $\mu_2 = \int w^2 K(w)dw$, and $v_j = \int w^j K^2(w)dw$ for j=0 and 2. Choosing $K(\cdot)$ to be a symmetric kernel, under certain mild conditions, we can follow the same arguments in the paper to show

$$\sqrt{\tilde{\mathbf{n}} h_2 h_1^3}\left[\begin{pmatrix} \tilde{\boldsymbol{\alpha}}(u) - \boldsymbol{\alpha}_0(u) \\ h_2(\tilde{\boldsymbol{b}}(u) - \boldsymbol{b}_0(u)) \end{pmatrix} - \tilde{\Gamma}(u)^{-1}\left(\frac{h_2^2}{2}\tilde{\Omega}_1 \boldsymbol{\alpha}_0^{(2)}(u) - \frac{h_1^2}{2}\tilde{\Lambda}\right)\right]$$
$$\xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \frac{\int t^2\phi^2(t)dt}{f_U(u)}\tilde{\Gamma}(u)^{-1}\tilde{\Omega}_2\tilde{\Gamma}(u)^{-1}\right),$$

where $\tilde{\Gamma}(u) = \mathbb{E}\left[\begin{pmatrix} \mathbf{X}\mathbf{X}^T g^{(2)}(0 \mid \mathbf{X}) & \mathbf{0} \\ \mathbf{0} & \mu_2 \mathbf{X}\mathbf{X}^T g^{(2)}(0 \mid \mathbf{X}) \end{pmatrix} \middle| U = u\right]$,

$\tilde{\Lambda} = \mathbb{E}\left[\begin{pmatrix} \mathbf{X} g^{(3)}(0 \mid \mathbf{X}) \\ \mathbf{0} \end{pmatrix} \middle| U = u\right]$, $\tilde{\Omega}_1 = \mathbb{E}\left[\begin{pmatrix} \mu_2 \mathbf{X}\mathbf{X}^T g^{(2)}(0 \mid \mathbf{X}) \\ \mathbf{0} \end{pmatrix} \middle| U = u\right]$,

and $\tilde{\Omega}_2 = \mathbb{E}\left[\begin{pmatrix} v_0 \mathbf{X}\mathbf{X}^T g(0 \mid \mathbf{X}) & \mathbf{0} \\ \mathbf{0} & v_2 \mathbf{X}\mathbf{X}^T g(0 \mid \mathbf{X}) \end{pmatrix} \middle| U = u\right]$.

The proposed *varying coefficient spatial modal regression* serves as a complement to the varying coefficient spatial mean regression introduced in Lu et al. (2009). To ensure a valid interpretation of the spatial nonstationar-

ity of the regression relationship, we could further develop a goodness-of-fit test to determine whether some coefficients truly vary over space.

S6-3 Unconditional Spatial Modal Regression

It is intriguing to extend conditional spatial modal regression to the unconditional counterpart. According to the law of iterated expectations, the estimates of coefficients from the conditional mean regression also represent the effects on the unconditional population average of the dependent variable. However, such a property is no longer guaranteed when transitioning from the mean to the mode. Unlike the mean case, it is not necessary for the conditional mode effect and unconditional mode effect to be equal. To investigate the impact of changes in the independent variables on the unconditional mode of the dependent variable, we broaden the scope by extending the nonspatial-unconditional quantile regression introduced in Firpo et al. (2009). This extension leads to the proposal of *unconditional spatial modal regression* grounded in the influence function and kernel density estimation.

Firpo et al. (2009) developed an unconditional quantile regression (UQR) model based on the influence function (IF) and recentered influence function (RIF). The IF is an analytical technique that assesses the impact of a particular observation on distributional statistics and is defined as

$$\mathrm{IF}(y; v(F)) = \lim_{\varepsilon \to 0} \left[v\left((1-\varepsilon)\cdot F + \varepsilon \cdot \delta_y\right) - v(F) \right] / \varepsilon, \ 0 \leq \varepsilon \leq 1,$$

where F represents the cumulative distribution function for Y and δ_y is a distribution that only puts mass at the value y. For a specific quantile τ of the outcome distribution, we can obtain

$$\mathrm{IF}\left(Y; q_\tau\right) = \left(\tau - \mathbb{I}\left\{Y \leq q_\tau\right\}\right) / f_Y\left(q_\tau\right),$$

where $f_Y(q_\tau)$ denotes the probability density function of Y evaluated at q_τ. Accordingly, we have $\mathbb{E}[\mathrm{IF}\left(Y; q_\tau\right)] = 0$. The RIF is defined as

$$\mathrm{RIF}\left(Y; q_\tau, F_Y\right) = q_\tau + \frac{\tau - \mathbb{I}\left\{Y \leq q_\tau\right\}}{f_Y\left(q_\tau\right)},$$

where q_τ is the value of the outcome variable Y at the quantile τ. According to Firpo et al. (2009), when we model the conditional expectation of $\mathrm{RIF}\left(Y; q_\tau, F_Y\right)$ as a function of explanatory variables, $\mathbb{E}(\mathrm{RIF}\left(Y; q_\tau, F_Y\right) \mid \mathbf{X} = \mathbf{x}) = m(\mathbf{x})$, a RIF regression can be viewed as an UQR. We then base the unconditional spatial modal regression on UQR such that

$$\mathbb{E}(\mathrm{RIF}\left(Y; q_{\tau^*}, F_Y\right) \mid \mathbf{X} = \mathbf{x}) = m^*(\mathbf{x}),$$

in which τ^* represents the τth quantile of Y evaluated at the mode value, and $m^*(\mathbf{x})$ is the unconditional spatial modal regression line. As a result, unconditional spatial modal regression can be regarded as a conventional regression with a transformed dependent variable.

S6-4 Spatial Modal Autoregressive Model

We in this paper investigate spatial modal regression under the presence of a

mixing condition. It is intriguing to extend the proposed modal regression framework to incorporate spatial dependence among cross-sectional units, which we refer to as the *spatial modal autoregressive model*. Following the notations in Su and Yang (2011), the model is expressed as follows

$$\mathbf{Y}_n = \lambda_0 \mathbf{W}_n \mathbf{Y}_n + \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{U}_n,$$

where n represents the total number of spatial units, $\mathbf{Y}_n \equiv (y_{n,1}, \cdots, y_{n,n})^T$ indicates a $n \times 1$ vector of response values, λ_0 is the spatial lag parameter, $\mathbf{W}_n \equiv \{w_{n,ij}\}$ is a known $n \times n$ spatial weight matrix with zero diagonal elements, $\mathbf{W}_n \mathbf{Y}_n$ is the spatial lagged variable, $\mathbf{X}_n \equiv \left(x_{n,1}^T, \cdots, x_{n,n}^T\right)^T$ is a $n \times p$ matrix containing the values of the regressors, $\boldsymbol{\beta}_0$ is a p-vector of regression coefficients, and $\mathbf{U}_n \equiv (u_{n,1}, \cdots, u_{n,n})^T$ denotes a n-vector of i.i.d. random disturbances with zero mode.

Because of the presence of the endogenous covariate (the spatial lag), the direct utilization of the kernel-based objective function in **S6-1** is not feasible. To address this, we propose an instrumental variable spatial modal regression estimation procedure, building on the framework introduced by Su and Yang (2011). Suppose that there is a $n \times q$ instrumental matrix $\mathbf{Z}_n \equiv (z_{n,1}, \cdots, z_{n,n})^T$ such that

$$Mode\left(y_{n,i} \leq \lambda_{0\tau} \bar{y}_{n,i} + \boldsymbol{\beta}_{0\tau}^T x_{n,i} \mid x_{n,i}, z_{n,i}\right) = 0, \ i = 1, \cdots, n.$$

Then, the following kernel-based objective function can be applied

$$Q_n(\lambda, \boldsymbol{\beta}, \boldsymbol{\gamma}) \equiv \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y_{n,i} - \lambda \bar{y}_{n,i} - \boldsymbol{\beta}^T x_{n,i} - \boldsymbol{\gamma}^T z_{n,i}}{h}\right),$$

where $\boldsymbol{\gamma}$ is the instrumental variable coefficient vector. Let $\xi_{n,i} \equiv (x_{n,i}^T, z_{n,i}^T)^T$. The estimation steps for obtaining spatial modal estimators are summarized as follows:

(i) for a given value of λ, perform a modal regression of $y_{n,i} - \lambda \bar{y}_{n,i}$ on $\xi_{n,i}$ to obtain $(\hat{\boldsymbol{\beta}}_n(\lambda), \hat{\boldsymbol{\gamma}}_n(\lambda)) \equiv \arg\max_{(\boldsymbol{\beta}, \boldsymbol{\gamma})} Q_n(\lambda, \boldsymbol{\beta}, \boldsymbol{\gamma})$;

(ii) minimize a weighted norm of $\hat{\boldsymbol{\gamma}}_n(\lambda)$ over λ to achieve the modal estimator of λ_0, i.e., $\hat{\lambda}_n = \arg\min_{\lambda} \hat{\boldsymbol{\gamma}}_n(\lambda)^T \hat{\boldsymbol{A}}_n \hat{\boldsymbol{\gamma}}_n(\lambda)$, where $\hat{\boldsymbol{A}}_n = \boldsymbol{A} + o_p(1)$ for some positive definite matrix $\boldsymbol{A}$;

(iii) run a modal regression of $y_{n,i} - \hat{\lambda}_{n\tau} \bar{y}_{n,i}$ on $\xi_{n,i}$ to obtain the modal estimator of $\boldsymbol{\beta}_0$, i.e., $\hat{\boldsymbol{\beta}}_n \equiv \hat{\boldsymbol{\beta}}_n(\hat{\lambda}_n)$.

This proposed estimation procedure not only addresses the spatial modal autoregressive model but also provides a solution to the endogeneity issue in modal regression. The detailed discussions, including asymptotic properties, will be explored in a separate paper.

S7 Technical Proofs

Due to the correlation structure inherent in spatial data, the asymptotic distributions of the modal estimators become intricate. The subsequent proofs heavily rely on the significance of the following lemmas, where we

use the notation $o_p(1)$ to denote a sequence of random variables that converges to zero in probability, i.e., $Z_n = o_p(1)$ implies $Z_n \xrightarrow{p} 0$ as $n \to \infty$.

Lemma 1 Suppose that $\alpha\left(\mathcal{B}(S), \mathcal{B}\left(S'\right)\right)$ in Definition 2.2 holds. Denote $\mathcal{L}_r(\mathcal{F})$ be the class of $\mathcal{F}$-measurable random variable $\mathbf{X}$ satisfying $\|\mathbf{X}\|_r = \left(\mathbb{E}|\mathbf{X}|^r\right)^{1/r} < \infty$. Suppose $\mathbf{X} \in \mathcal{L}_r(\mathcal{B}(S))$ and $\mathbf{Y} \in \mathcal{L}_s\left(\mathcal{B}\left(S'\right)\right)$. Assume also that $1 \leq r, s, t < \infty$ and $r^{-1} + s^{-1} + t^{-1} = 1$. Then,

$$|\mathbb{E}\mathbf{X}\mathbf{Y} - \mathbb{E}\mathbf{X}\mathbb{E}\mathbf{Y}| \leq C\|\mathbf{X}\|_r\|\mathbf{Y}\|_s \left\{\chi\left(\text{Card}(S), \text{Card}\left(S'\right)\right)\varphi\left(\text{dist}\left(S, S'\right)\right)\right\}^{1/t}.$$

For random variables bounded with probability one, the right-hand side of the above equation can be replaced by $C\chi\left(\text{Card}(S), \text{Card}\left(S'\right)\right)\varphi\left(\text{dist}\left(S, S'\right)\right)$.

Proof. The proof can be seen in Tran (1990).

□

Lemma 2 Let $\phi(\cdot)$ be a standard Gaussian kernel, $S(\mathbf{X_i}) = m(\mathbf{X_i}) - m(\mathbf{x}) - m^{(1)}(\mathbf{x})(\mathbf{X_i} - \mathbf{x})$, and $\mu_2 = \int_{\mathbb{R}^d} \mathbf{u}\mathbf{u}^T K(\mathbf{u}) d\mathbf{u}$. We then have the following equations

$$\frac{1}{\tilde{\mathbf{n}} h_1^3 h_2^d} \sum_{\mathbf{i} \in I_\mathbf{n}} \phi^{(2)}\left(\frac{\varepsilon_\mathbf{i}}{h_1}\right)\left(\frac{\mathbf{X_i} - \mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X_i} - \mathbf{x}}{h_2}\right)^T K\left(\frac{\mathbf{X_i} - \mathbf{x}}{h_2}\right)$$

$$= f(\mathbf{x})\mu_2 g^{(2)}(0 \mid \mathbf{X} = \mathbf{x}) + o_p(1).$$

$$\frac{1}{\tilde{\mathbf{n}} h_1^3 h_2^d} \sum_{\mathbf{i} \in I_\mathbf{n}} \phi^{(2)}\left(\frac{\varepsilon_\mathbf{i}}{h_1}\right) K\left(\frac{\mathbf{X_i} - \mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X_i} - \mathbf{x}}{h_2}\right) S(\mathbf{X_i})$$

$$= \frac{h_2^2}{2} f(\mathbf{x}) g^{(2)}(0 \mid \mathbf{X} = \mathbf{x}) \sum_{i=1}^{d}\sum_{j=1}^{d} m_{ij}(\mathbf{x}) \int_{\mathbb{R}^d} u_i u_j \mathbf{u} K(\mathbf{u}) d\mathbf{u} + o_p(1).$$

$$\frac{1}{\tilde{\mathbf{n}}h_1^2h_2^d}\sum_{\mathbf{i}\in I_\mathbf{n}}\phi^{(1)}\left(\frac{\varepsilon_\mathbf{i}}{h_1}\right)K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)=\frac{h_1^2}{2}f(\mathbf{x})g^{(3)}(0\mid\mathbf{X}=\mathbf{x})+o_p(1).$$

Proof. We prove the first equation. The second and third equations can be proved by following the same steps. Let $T_\mathbf{i}=\frac{1}{h_1^3h_2^d}\phi^{(2)}\left(\frac{\varepsilon_\mathbf{i}}{h_1}\right)\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)^T K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)$. When $\mathbf{n}\to\infty$, we have

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{\tilde{\mathbf{n}}}\sum_{\mathbf{i}\in I_\mathbf{n}}T_\mathbf{i}\right)=\mathbb{E}(\mathbb{E}(T_\mathbf{i}\mid\mathbf{X_i}))\\
=&\frac{1}{h_1^3h_2^d}\iint_{\mathbb{R}^{d+1}}\phi^{(2)}\left(\frac{\varepsilon}{h_1}\right)\left(\frac{\mathbf{X}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X}-\mathbf{x}}{h_2}\right)^T K\left(\frac{\mathbf{X}-\mathbf{x}}{h_2}\right)g(\varepsilon|\mathbf{X})d\varepsilon dF(\mathbf{X})\\
=&\frac{1}{h_1^2}\iint_{\mathbb{R}^{d+1}}\phi(\tau)(\tau^2-1)\mathbf{u}\mathbf{u}^TK(\mathbf{u})g(\tau h_1\mid\mathbf{X}=\mathbf{x})f(\mathbf{u}h_2+\mathbf{x})d\tau d\mathbf{u}\\
=&f(\mathbf{x})\mu_2g^{(2)}(0\mid\mathbf{X}=\mathbf{x}).
\end{aligned}$$

We first consider the variance of the above equation, where

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{\tilde{\mathbf{n}}}\sum_{\mathbf{i}\in I_\mathbf{n}}T_\mathbf{i}\right)&=\frac{1}{\tilde{\mathbf{n}}^2}\sum_{\mathbf{i}\in I_\mathbf{n}}\mathrm{Var}(T_\mathbf{i})+\frac{1}{\tilde{\mathbf{n}}^2}\sum_{\mathbf{i}\neq\mathbf{j}}\mathbb{E}(T_\mathbf{i}-\mathbb{E}(T_\mathbf{i}))(T_\mathbf{j}-\mathbb{E}(T_\mathbf{j}))^T\\
&=:A_1+A_2.
\end{aligned}$$

With the assumption that $\tilde{\mathbf{n}}h_2^dh_1^5\to\infty$ and the Lebesgue dominated convergence theorem, we have

$$\begin{aligned}
A_1\le&\frac{1}{\tilde{\mathbf{n}}^2}\sum_{\mathbf{i}\in I_\mathbf{n}}\mathbb{E}(T_\mathbf{i}T_\mathbf{i}^T)=\frac{1}{h_1^6h_2^{2d}}\iint_{\mathbb{R}^{d+1}}\left(\phi^{(2)}\left(\frac{\varepsilon}{h_1}\right)\right)^2\left(\frac{\mathbf{X}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X}-\mathbf{x}}{h_2}\right)^T\\
&K^2\left(\frac{\mathbf{X}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X}-\mathbf{x}}{h_2}\right)^T g(\varepsilon\mid\mathbf{X})d\varepsilon dF(\mathbf{X})\\
=&\frac{1}{\tilde{\mathbf{n}}h_2^dh_1^5}\mathbb{E}[(\phi^{(2)}(\tau))^2\mid\mathbf{X}=\mathbf{x}]f(\mathbf{x})\int_{\mathbb{R}}\mathbf{u}\mathbf{u}^TK^2(\mathbf{u})(\mathbf{u}\mathbf{u}^T)^Td\mathbf{u}+o\left(\frac{1}{\tilde{\mathbf{n}}h_2^dh_1^5}\right)\to 0.
\end{aligned}$$

We now prove that for $\mathbf{n}$ large enough, there exists C such that $\tilde{\mathbf{n}}h_2^d h_1^5 A_2 < C$. Define $S = \{\mathbf{i}, \mathbf{j}, d(\mathbf{i}, \mathbf{j}) \leq s_{\mathbf{n}}\}$, where $s_{\mathbf{n}}$ is a real sequence that converges to infinity and will be specified later. We have $A_2 = A_{21} + A_{22}$ with

$$A_{21} = \frac{1}{\tilde{\mathbf{n}}^2} \sum_{\mathbf{i},\mathbf{j} \in S} \mathbb{E}(T_{\mathbf{i}} - \mathbb{E}(T_{\mathbf{i}}))(T_{\mathbf{j}} - \mathbb{E}(T_{\mathbf{j}}))^T,$$

$$A_{22} = \frac{1}{\tilde{\mathbf{n}}^2} \sum_{\mathbf{i},\mathbf{j} \in S^c} \mathbb{E}(T_{\mathbf{i}} - \mathbb{E}(T_{\mathbf{i}}))(T_{\mathbf{j}} - \mathbb{E}(T_{\mathbf{j}}))^T,$$

where S^c stands for the complement of S. In addition, we have

$$\frac{1}{h_1^4}\mathbb{E}\left[\phi^{(2)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)\phi^{(2)}\left(\frac{\varepsilon_{\mathbf{j}}}{h_1}\right)\Big|(\mathbf{X_i}, \mathbf{X_j})\right] = \frac{1}{h_1^2}\iiint_{\mathbb{R}^{2d+2}} \phi(t)(t^2-1)\phi(s)$$

$$(s^2-1)g(th_1, sh_1|(\mathbf{X_i}, \mathbf{X_j}))dtdsdF(\mathbf{X_i}, \mathbf{X_j}) = O(1).$$

Similarly, we obtain $\mathbb{E}\left[\frac{1}{h_1^2}\phi^{(2)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)\Big|\mathbf{X_i}\right]^2 = O(h_1^{-3})$. In addition, by condition C3, we have

$$\mathbb{E}\Bigg[\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)^T K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)$$

$$\left(\left(\frac{\mathbf{X_j}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X_j}-\mathbf{x}}{h_2}\right)^T K\left(\frac{\mathbf{X_j}-\mathbf{x}}{h_2}\right)\right)^T\Bigg]$$

$$= h_2^{2d}\iint_{\mathbb{R}^{2d}} \mathbf{u}\mathbf{u}^T K(\mathbf{u})(\mathbf{v}\mathbf{v}^T K(\mathbf{v}))^T f(\mathbf{u}h_2 + \mathbf{x}, \mathbf{v}h_2 + \mathbf{x})d\mathbf{u}d\mathbf{v} = O(h_2^{2d}).$$

Also, we can get

$$\mathbb{E}\left[\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)^T K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\right]$$

$$= h_2^d \int_{\mathbb{R}^d} \mathbf{u}\mathbf{u}^T K(\mathbf{u}) f(\mathbf{u}h_2 + \mathbf{x})d\mathbf{u} = O(h_2^d).$$

Let us consider A_{21}, where we can obtain

$$\left|\frac{1}{\tilde{\mathbf{n}}^2}\mathbb{E}(T_{\mathbf{i}}-\mathbb{E}(T_{\mathbf{i}}))(T_{\mathbf{j}}-\mathbb{E}(T_{\mathbf{j}}))^T\right| = \left|\frac{1}{\tilde{\mathbf{n}}^2}(\mathbb{E}(T_{\mathbf{i}}T_{\mathbf{j}}^T)-\mathbb{E}(T_{\mathbf{i}})\mathbb{E}(T_{\mathbf{j}})^T)\right|$$

$$\leq \frac{1}{\tilde{\mathbf{n}}^2}\mathbb{E}[\mathbb{E}|T_{\mathbf{i}}T_{\mathbf{j}}^T|(\mathbf{X}_{\mathbf{i}},\mathbf{X}_{\mathbf{j}})|] + \frac{1}{\tilde{\mathbf{n}}^2}(\mathbb{E}[\mathbb{E}|T_{\mathbf{i}}| \mid \mathbf{X}_{\mathbf{i}}])(\mathbb{E}[\mathbb{E}|T_{\mathbf{i}}| \mid \mathbf{X}_{\mathbf{i}}])^T \leq \tilde{\mathbf{n}}^{-2}h_2^{-2d}h_1^{-6}$$

$$\mathbb{E}\left[\left(\frac{\mathbf{X}_{\mathbf{i}}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X}_{\mathbf{i}}-\mathbf{x}}{h_2}\right)^T K\left(\frac{\mathbf{X}_{\mathbf{i}}-\mathbf{x}}{h_2}\right)\left(\left(\frac{\mathbf{X}_{\mathbf{j}}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X}_{\mathbf{j}}-\mathbf{x}}{h_2}\right)^T\right.\right.$$

$$\left.\left.K\left(\frac{\mathbf{X}_{\mathbf{j}}-\mathbf{x}}{h_2}\right)\right)^T\right]\mathbb{E}\left[\phi^{(2)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)\phi^{(2)}\left(\frac{\varepsilon_{\mathbf{j}}}{h_1}\right)\Big|(\mathbf{X}_{\mathbf{i}},\mathbf{X}_{\mathbf{j}})\right] + \tilde{\mathbf{n}}^{-2}h_2^{-2d}h_1^{-6}$$

$$\left(\mathbb{E}\left[\left(\frac{\mathbf{X}_{\mathbf{i}}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X}_{\mathbf{i}}-\mathbf{x}}{h_2}\right)^T K\left(\frac{\mathbf{X}_{\mathbf{i}}-\mathbf{x}}{h_2}\right)\right]\mathbb{E}\left[\phi^{(2)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)\Big|\mathbf{X}_{\mathbf{i}}\right]\right)$$

$$\left(\mathbb{E}\left[\left(\frac{\mathbf{X}_{\mathbf{i}}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X}_{\mathbf{i}}-\mathbf{x}}{h_2}\right)^T K\left(\frac{\mathbf{X}_{\mathbf{i}}-\mathbf{x}}{h_2}\right)\right]\mathbb{E}\left[\phi^{(2)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)\Big|\mathbf{X}_{\mathbf{i}}\right]\right)^T$$

$$\leq C\tilde{\mathbf{n}}^{-2}h_1^{-5}h_2^{-d} \to 0.$$

After that, we compute A_{22}. Since kernel functions are bounded, by applying Lemma 1, setting $r = s = 2+\delta$ and $t = (2+\delta)/\delta$, we can get

$$\left|\frac{1}{\tilde{\mathbf{n}}^2}\mathbb{E}(T_{\mathbf{i}}-\mathbb{E}(T_{\mathbf{i}}))(T_{\mathbf{j}}-\mathbb{E}(T_{\mathbf{j}}))^T\right|$$

$$\leq C\tilde{\mathbf{n}}^{-2}h_2^{-2d}h_1^{-6}(h_1h_2^d)^{2/(2+\delta)}[\chi(1,1)\varphi(\|\mathbf{i}-\mathbf{j}\|)]^{\delta/(2+\delta)}.$$

Then, we obtain

$$\tilde{\mathbf{n}}h_2^dh_1^5\left|\frac{1}{\tilde{\mathbf{n}}^2}\sum_{\mathbf{i},\mathbf{j}\in S^c}\mathbb{E}(T_{\mathbf{i}}-\mathbb{E}(T_{\mathbf{i}}))(T_{\mathbf{j}}-\mathbb{E}(T_{\mathbf{j}}))^T\right| \leq C\tilde{\mathbf{n}}^{-1}h_2^{-d\delta/(2+\delta)}h_1^{-\delta/(2+\delta)}$$

$$\sum_{\mathbf{i},\mathbf{j}\in S^c}[\chi(1,1)\varphi(\|\mathbf{i}-\mathbf{j}\|)]^{\delta/(2+\delta)} \leq Ch_2^{-d\delta/(2+\delta)}h_1^{-\delta/(2+\delta)}s_{\mathbf{n}}^{-N\delta/(2+\delta)}$$

$$\sum_{\|\mathbf{i}\|>s_{\mathbf{n}}}[\|\mathbf{i}\|^N\varphi(\|\mathbf{i}\|)]^{\delta/(2+\delta)} \leq Ch_2^{-d\delta/(2+\delta)}h_1^{-\delta/(2+\delta)}s_{\mathbf{n}}^{-N\delta/(2+\delta)}\sum_{\|\mathbf{i}\|>s_{\mathbf{n}}}[\|\mathbf{i}\|^{N-\mu}]^{\delta/(2+\delta)}.$$

As $\mu > N+1$, we choose $s_{\mathbf{n}} = (h_2^dh_1)^{-1/N}$, which gives

$$\mathrm{Var}\left(\frac{1}{\tilde{\mathbf{n}}}\sum_{\mathbf{i}\in I_{\mathbf{n}}} T_{\mathbf{i}}\right) = O\left(\frac{1}{\tilde{\mathbf{n}}h_2^d h_1^5}\right).$$

Combining the above calculations, we achieve the desired result.

□

The second lemma holds independent significance and plays a pivotal role in the subsequent proofs. In alignment with Hallin et al. (2009), we refrain from assuming that the mixing coefficient adheres to the form specified in Definition 2.2 in the paper.

Lemma 3 Let $\left\{(Y_{\mathbf{j}}, \mathbf{X}_{\mathbf{j}})\,; \mathbf{j} \in \mathbb{Z}^N\right\}$ be a stationary spatial process with general mixing coefficient $\varphi(\mathbf{j}) = \varphi(j_1, \cdots, j_N) := \sup\{|\mathrm{P}(AB) - \mathrm{P}(A)\mathrm{P}(B)| : A \in \mathcal{B}(\{Y_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}\}), B \in \mathcal{B}(\{Y_{\mathbf{i}+\mathbf{j}}, \mathbf{X}_{\mathbf{i}+\mathbf{j}}\})\}$. Define $A_{\mathbf{n}} = \left(\frac{1}{\tilde{\mathbf{n}}h_1h_2^d}\right)^{1/2} \sum_{\mathbf{i}\in I_{\mathbf{n}}} \left[\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right) K\left(\frac{\mathbf{X}_{\mathbf{i}}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X}_{\mathbf{i}}-\mathbf{x}}{h_2}\right) - \mathbb{E}\left(\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right) K\left(\frac{\mathbf{X}_{\mathbf{i}}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X}_{\mathbf{i}}-\mathbf{x}}{h_2}\right)\right)\right] = (\tilde{\mathbf{n}})^{-1/2} (h_1h_2^d)^{-1/2} \sum_{\mathbf{i}\in I_{\mathbf{n}}} \Delta_{\mathbf{j}}(\mathbf{x})$ and $\mathrm{Var}(A_{\mathbf{n}}) = (\tilde{\mathbf{n}}h_1h_2^d)^{-1} \sum_{\mathbf{i}\in I_{\mathbf{n}}} \mathbb{E}\left[\Delta_{\mathbf{j}}^2(\mathbf{x})\right] + (\tilde{\mathbf{n}})^{-1/2} (h_1h_2^d)^{-1} \sum_{\{\mathbf{i},\mathbf{j}\in I_{\mathbf{n}} | \exists k: i_k \neq j_k\}} \mathbb{E}\left[\Delta_{\mathbf{i}}(\mathbf{x})\Delta_{\mathbf{j}}(\mathbf{x})\right] = \tilde{I}(\mathbf{x}) + \tilde{R}(\mathbf{x})$, where $\Delta_{\mathbf{i}}(\mathbf{x}) = \phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right) K\left(\frac{\mathbf{X}_{\mathbf{i}}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X}_{\mathbf{i}}-\mathbf{x}}{h_2}\right)$. Then, for any $\mathbf{c}_{\mathbf{n}} = (c_{\mathrm{n}1}, \cdots, c_{\mathrm{n}N}) \in \mathbb{Z}^N$ with $1 < c_{\mathrm{n}k} < n_k$ for all $k = 1, \cdots, N$, we achieve

$$|\tilde{R}(\mathbf{x})| \leq C\left(\tilde{\mathbf{n}}h_1h_2^d\right)^{-1}\left[\tilde{J}_1(\mathbf{x}) + \tilde{J}_2(\mathbf{x})\right] \to 0,$$

where we have $\tilde{J}_1(\mathbf{x}) = h_1^5 h_2^{2d} \prod_{k=1}^N (n_k c_{\mathbf{n}k})$ and $\tilde{J}_2(\mathbf{x}) = (h_1h_2^d)^{2/(2+\delta)} \tilde{\mathbf{n}} \sum_{k=1}^N \left(\sum_{|j_s|=1, s=1,\cdots,k-1}^{n_s} \sum_{|j_k|=c_{\mathbf{n}\mathbf{k}}}^{n_k} \sum_{|j_s|=1, s=k+1,\cdots,N}^{n_s} \{\varphi(j_1, \cdots, j_N)\}^{\delta/(2+\delta)}\right)$. If furthermore $\varphi(j_1, \cdots, j_N)$ takes the form $\varphi(\|\mathbf{j}\|)$, we have $\tilde{J}_2(\mathbf{x}) \leq C(h_1h_2^d)^{2/(2+\delta)}$

$\tilde{\mathbf{n}} \sum_{k=1}^{N} \left(\sum_{t=c_{\mathbf{n}k}}^{\|\mathbf{n}\|} t^{N-1} \{\varphi(t)\}^{\delta/(2+\delta)} \right)$.

Proof. According to the assumptions in this paper and the Lebesgue density theorem, we obtain

$$
\begin{aligned}
&(h_1 h_2^d)^{-1} \mathbb{E}[\Delta_{\mathbf{i}}(\mathbf{x}) \Delta_{\mathbf{j}}(\mathbf{x})] = (h_1 h_2^d)^{-1} \Big\{ \mathbb{E}\Big[\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right) K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right) \\
&\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{j}}}{h_1}\right) K\left(\frac{\mathbf{X_j}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X_j}-\mathbf{x}}{h_2}\right) \Big] - \mathbb{E}\Big[\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right) K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right) \\
&\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right) \Big] \mathbb{E}\Big[\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{j}}}{h_1}\right) K\left(\frac{\mathbf{X_j}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X_j}-\mathbf{x}}{h_2}\right) \Big] \Big\} \\
=& h_1 h_2^d \iiiint \phi(t)\, tK(\mathbf{u})\mathbf{u}\phi(s)\, sK(\mathbf{v})\mathbf{v} g(th_1, sh_1 \mid \mathbf{x_i}, \mathbf{x_j}) \\
&f(\mathbf{u}h_2 + \mathbf{x}, \mathbf{v}h_2 + \mathbf{x}) dt d\mathbf{u} ds d\mathbf{v} - h_1 h_2^d \iint \phi(t)\, tg(th_1 \mid \mathbf{x}) K(\mathbf{u})\mathbf{u} \\
&f(uh_2 + \mathbf{x}) dt d\mathbf{u} \iint \phi(s)\, sg(sh_1 \mid \mathbf{x}) K(\mathbf{v})\mathbf{v} f(vh_2 + \mathbf{x}) ds d\mathbf{v} \leq C h_1^4 h_2^d.
\end{aligned}
$$

Let $c_{\mathbf{n}} = (c_{\mathbf{n}1}, \cdots, c_{\mathbf{n}N}) \in \mathbb{R}^N$ be a sequence of vectors with positive components. Define

$$\mathcal{S}_1 := \{\mathbf{i} \neq \mathbf{j} \in I_{\mathbf{n}} : |j_k - i_k| \leq c_{\mathbf{n}k}, \text{ for all } k = 1, \cdots, N\},$$

$$\mathcal{S}_2 := \{\mathbf{i}, \mathbf{j} \in I_{\mathbf{n}} : |j_k - i_k| > c_{\mathbf{n}k}, \text{ for some } k = 1, \cdots, N\}.$$

Clearly,

$$\text{Card}\,(\mathcal{S}_1) \leq 2^N \tilde{\mathbf{n}} \prod_{k=1}^{N} c_{\mathbf{n}k}.$$

Splitting $\tilde{R}(\mathbf{x})$ into $\left(\tilde{\mathbf{n}} h_1 h_2^d\right)^{-1} (J_1 + J_2)$, with $J_\ell := \sum_{\mathbf{i},\mathbf{j} \in \mathcal{S}_\ell} \mathbb{E}\Delta_{\mathbf{j}}(\mathbf{x})\Delta_{\mathbf{i}}(\mathbf{x}), \ell = 1, 2$, it follows from the above calculation that

$$|J_1| \leq C h_1^5 h_2^{2d}\, \text{Card}\,(\mathcal{S}_1) \leq 2^N C h_1^5 h_2^{2d} \tilde{\mathbf{n}} \prod_{k=1}^{N} c_{\mathbf{n}k}.$$

Turning to J_2, we have $|J_2| \leq \sum_{i,j\in\mathcal{S}_2}|\mathbb{E}\Delta_{\mathbf{j}}(\mathbf{x})\Delta_{\mathbf{i}}(\mathbf{x})|$. According to Lemma 1 and the boundedness of $\Delta_{\mathbf{i}}(\mathbf{x})$, setting $r = s = 2+\delta$ and $t = (2+\delta)/\delta$ yields

$$
\begin{aligned}
&|\mathbb{E}\Delta_{\mathbf{j}}(\mathbf{x})\Delta_{\mathbf{i}}(\mathbf{x})| \\
\leq& C\left(\mathbb{E}\Big[\Big|\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\Big|^{2+\delta}\Big]\right)^{2/(2+\delta)}\{\varphi(\mathbf{j}-\mathbf{i})\}^{\delta/(2+\delta)} \\
\leq& C(h_1h_2^d)^{2/(2+\delta)}\left((h_1^{-1}h_2^{-d})\mathbb{E}\Big[\Big|\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\Big|^{2+\delta}\Big]^{2/(2+\delta)}\right) \\
&\{\varphi(\mathbf{j}-\mathbf{i})\}^{\delta/(2+\delta)} \leq C(h_1h_2^d)^{2/(2+\delta)}\{\varphi(\mathbf{j}-\mathbf{i})\}^{\delta/(2+\delta)}.
\end{aligned}
$$

Hence,

$$|J_2| \leq C(h_1h_2^d)^{2/(2+\delta)}\sum\sum_{\mathbf{i},\mathbf{j}\in\mathcal{S}_2}\{\varphi(\mathbf{j}-\mathbf{i})\}^{\delta/(2+\delta)}.$$

For any N-tuple $\mathbf{0} \neq \ell = (\ell_1,\cdots,\ell_N) \in \{0,1\}^N$, set

$$
\begin{aligned}
\mathcal{S}(\ell_1,\cdots,\ell_N) := \{&\mathbf{i},\mathbf{j}\in I_{\mathbf{n}} : |j_k - i_k| > c_{\mathbf{n}k} \\
&\text{if } \ell_k = 1 \text{ and } |j_k - i_k| \leq c_{\mathbf{n}k} \text{ if } \ \ell_k = 0, k = 1,\cdots,N\}
\end{aligned}
$$

and

$$V(\ell_1,\cdots,\ell_N) := \sum\sum_{\mathbf{i},\mathbf{j}\in\mathcal{S}(\ell_1,\cdots,\ell_N)}\{\varphi(\mathbf{j}-\mathbf{i})\}^{\delta/(2+\delta)}.$$

Then, we can get

$$
\begin{aligned}
\sum\sum_{\mathbf{i},\mathbf{j}\in\mathcal{S}_2}\{\varphi(\mathbf{j}-\mathbf{i})\}^{\delta/(2+\delta)} &= \sum_{\mathbf{0}\neq\ell\in\{0,1\}^N} V(\ell_1,\cdots,\ell_N) \\
&\leq \sum_{\mathbf{0}\neq\ell\in\{0,1\}^N}\tilde{\mathbf{n}}\sum_{|j_1|}\cdots\sum_{|j_k|}\cdots\sum_{|j_N|}\{\varphi(j_1,\cdots,j_N)\}^{\delta/(2+\delta)}
\end{aligned}
$$

with the sums $\sum_{|j_k|}$ running over all j_k's such that $1 \leq |j_k| \leq n_k$ when $\ell_k = 0$, and $c_{\mathbf{n}1} \leq |j_k| \leq n_k$ when $\ell_k = 1$. Since all terms are non-negative, for $1 \leq c_{\mathbf{n}k} \leq n_k$, sums of the form $\sum_{|j_k|=c_{\mathbf{n}k}}^{n_k} \cdots$ are smaller than those of the form $\sum_{|j_k|=1}^{n_k} \cdots$. Thus, following the same arguments as Hallin et al. (2009), we can achieve

$$|J_2| \leq C(h_1 h_2^d)^{2/(2+\delta)} \tilde{\mathbf{n}} \sum_{k=1}^{N} \sum_{|j_1|=1}^{n_1} \cdots \sum_{|j_{k-1}|=1}^{n_{k-1}} \sum_{|j_k|=c_{\mathbf{n}k}}^{n_k} \sum_{|j_{k+1}|=1}^{n_{k+1}} \cdots \sum_{|j_N|=1}^{n_N} \{\varphi(j_1, \cdots, j_N)\}^{\delta/(2+\delta)}.$$

If $\varphi(j_1, \cdots, j_N)$ depends on $\|\mathbf{j}\|$ only, we obtain

$$\begin{aligned} &\sum_{|j_1|=1}^{n_1} \cdots \sum_{|j_{k-1}|=1}^{n_{k-1}} \sum_{|j_k|=c_{\mathbf{n}k}}^{n_k} \sum_{|j_{k+1}|=1}^{n_{k+1}} \cdots \sum_{|j_N|=1}^{n_N} \{\varphi(\|\mathbf{j}\|)\}^{\delta/(2+\delta)} \\ \leq &\sum_{|j_1|=1}^{\|\mathbf{n}\|} \cdots \sum_{|j_{N-1}|=1}^{t} \{\varphi(t)\}^{\delta/(2+\delta)} \leq \sum_{t=c_{\mathbf{n}k}}^{\|\mathbf{n}\|} t^{N-1} \{\varphi(t)\}^{\delta/(2+\delta)}. \end{aligned}$$

After that, we need to show $|\tilde{R}(\mathbf{x})| \to 0$. Let $c_{\mathbf{n}k}^a := (h_1 h_2^d)^{-\delta/(2+\delta)} \to \infty$. Clearly, $c_{\mathbf{n}k} < n_k$ because of $n_k (h_1 h_2^d)^{\delta/(2+\delta)a} > 1$ for all k. Based on the above results, with $a > \delta N/(2+\delta)$, we arrive at

$$\left(\tilde{\mathbf{n}} h_1 h_2^d\right)^{-1} J_2 \leq C \sum_{k=1}^{N} \left(c_{\mathbf{n}k}^a \sum_{t=c_{\mathbf{n}k}}^{\infty} t^{N-1} \{\varphi(t)\}^{\delta/(2+\delta)} \right) \to 0,$$

$$\left(\tilde{\mathbf{n}} h_1 h_2^d\right)^{-1} J_1 \leq C h_1^4 h_2^d c_{\mathbf{n}1} \cdots c_{\mathbf{n}N} = C h_1^4 h_2^d (h_1 h_2^d)^{-\delta N/(2+\delta)a} \to 0.$$

□

Lemma 4 Let the spatial process $\{Y_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}\}$ satisfy the mixing property in

Definition 2.2, and denote $\tilde{U}_j, j = 1, \cdots, M$, as a $M-$tuple of measurable functions such that $\tilde{U}_j$ is measurable with respect to $\left\{\left(Y_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}\right), \mathbf{i} \in \tilde{\ell}_j\right\}$, where $\tilde{l}_j \subset \ell_{\mathbf{n}}$. If $\text{Card}\left(\tilde{\ell}_j\right) \leq p$ and $d\left(\tilde{\ell}_\ell, \tilde{\ell}_j\right) \geq q$ for any $\ell \neq j$, then

$$\left| \mathbb{E}\left[\exp\left\{iu\sum_{j=1}^{M}\tilde{U}_j\right\}\right] - \prod_{j=1}^{M}\mathbb{E}\left[\exp\left\{iu\tilde{U}_j\right\}\right] \right| \leq C\sum_{j=1}^{M-1}\chi(p,(M-j)p)\varphi(q),$$

where $i = \sqrt{-1}$.

Proof. The proof can be seen in Hallin et al. (2009).

□

S7-1 Proof of Theorem 2.1

Following Ullah et al. (2023), let $\alpha_{\mathbf{n}} = (\tilde{\mathbf{n}}h_1^3h_2^d)^{-1/2} + h_1^2 + h_2^2$. We need to show that for any given $\eta > 0$, there exists a constant c such that

$$\mathrm{P}\left\{\sup_{\|\boldsymbol{\mu}\|=c} Q_{\tilde{\mathbf{n}}}\left(\boldsymbol{\theta} + \alpha_{\mathbf{n}}\boldsymbol{\mu}\right) < Q_{\tilde{\mathbf{n}}}\left(\boldsymbol{\theta}\right)\right\} \geqslant 1 - \eta.$$

By rewriting the kernel-based objective function, we have

$$Q_{\tilde{\mathbf{n}}}(\boldsymbol{\theta}) = \frac{1}{\tilde{\mathbf{n}}h_1h_2^d}\sum_{\mathbf{i}\in I_{\mathbf{n}}}\phi\left(\frac{Y_{\mathbf{i}} - \boldsymbol{\theta}^T\mathbf{X}_{\mathbf{i}}^*}{h_1}\right)K\left(\frac{\mathbf{X}_{\mathbf{i}} - \mathbf{x}}{h_2}\right),$$

where $\mathbf{X}_{\mathbf{i}}^* = [1\ (\mathbf{X}_{\mathbf{i}} - \mathbf{x})h_2^{-1}]^T$ is the rescaled local design vector. To study the local perturbation $\boldsymbol{\theta} \to \boldsymbol{\theta} + \alpha_{\mathbf{n}}\boldsymbol{\mu}$, we use the following decomposition of the residuals, i.e., $Y_{\mathbf{i}} - \boldsymbol{\theta}^T\mathbf{X}_{\mathbf{i}}^* = \varepsilon_{\mathbf{i}} + S(\mathbf{X}_{\mathbf{i}})$ and $Y_{\mathbf{i}} - (\boldsymbol{\theta} + \alpha_{\mathbf{n}}\boldsymbol{\mu})^T\mathbf{X}_{\mathbf{i}}^* = \varepsilon_{\mathbf{i}} + S(\mathbf{X}_{\mathbf{i}}) - \alpha_{\mathbf{n}}\boldsymbol{\mu}^T\mathbf{X}_{\mathbf{i}}^*$, where $\varepsilon_{\mathbf{i}} = Y_{\mathbf{i}} - m(\mathbf{X}_{\mathbf{i}})$, and $S(\mathbf{X}_{\mathbf{i}}) = m(\mathbf{X}_{\mathbf{i}}) - m(\mathbf{x}) - m^{(1)}(\mathbf{x})^T(\mathbf{X}_{\mathbf{i}} - \mathbf{x})$ denotes the local linear approximation error. Using this

representation, the difference in the kernel-weighted objective function can be rewritten as

$$\begin{aligned}
&Q_{\tilde{\mathbf{n}}}\left(\boldsymbol{\theta}+\alpha_{\mathbf{n}}\boldsymbol{\mu}\right)-Q_{\tilde{\mathbf{n}}}\left(\boldsymbol{\theta}\right)\\
=&\frac{1}{\tilde{\mathbf{n}}h_1h_2^d}\sum_{\mathbf{i}\in I_{\mathbf{n}}}\phi\left(\frac{\varepsilon_{\mathbf{i}}+S(\mathbf{X_i})-\alpha_{\mathbf{n}}\boldsymbol{\mu}^T\mathbf{X}_{\mathbf{i}}^*}{h_1}\right)K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\\
&-\frac{1}{\tilde{\mathbf{n}}h_1h_2^d}\sum_{\mathbf{i}\in I_{\mathbf{n}}}\phi\left(\frac{\varepsilon_{\mathbf{i}}+S(\mathbf{X_i})}{h_1}\right)K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right).
\end{aligned}$$

Because $\alpha_{\mathbf{n}}\to 0$, the perturbation term $-\alpha_{\mathbf{n}}\boldsymbol{\mu}^T\mathbf{X}_{\mathbf{i}}^*$ is infinitesimal. Therefore, it is legitimate to invoke Taylor's theorem to expand $\phi(\cdot)$ around the unperturbed argument $(\varepsilon_{\mathbf{i}}+S(\mathbf{X_i}))/h_1$. Then, according to Taylor expansion, we can obtain

$$\begin{aligned}
&Q_{\tilde{\mathbf{n}}}\left(\boldsymbol{\theta}+\alpha_{\mathbf{n}}\boldsymbol{\mu}\right)-Q_{\tilde{\mathbf{n}}}\left(\boldsymbol{\theta}\right)\\
=&\frac{1}{\tilde{\mathbf{n}}h_1h_2^d}\sum_{\mathbf{i}\in I_{\mathbf{n}}}\Bigg[-\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}+S(\mathbf{X_i})}{h_1}\right)\left(\frac{\alpha_{\mathbf{n}}\boldsymbol{\mu}^T\mathbf{X}_{\mathbf{i}}^*}{h_1}\right)K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\\
&+\frac{1}{2}\phi^{(2)}\left(\frac{\varepsilon_{\mathbf{i}}+S(\mathbf{X_i})}{h_1}\right)\left(\frac{\alpha_{\mathbf{n}}\boldsymbol{\mu}^T\mathbf{X}_{\mathbf{i}}^*}{h_1}\right)^2K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\\
&-\frac{1}{6}\phi^{(3)}\left(\frac{\varepsilon_{\mathbf{i}}^*}{h_1}\right)\left(\frac{\alpha_{\mathbf{n}}\boldsymbol{\mu}^T\mathbf{X}_{\mathbf{i}}^*}{h_1}\right)^3K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\Bigg]=I_1+I_2+I_3,
\end{aligned}$$

where $\varepsilon_{\mathbf{i}}^*$ is between $\varepsilon_{\mathbf{i}}+S(\mathbf{X_i})$ and $\varepsilon_{\mathbf{i}}+S(\mathbf{X_i})-\alpha_{\mathbf{n}}\boldsymbol{\mu}^T\mathbf{X}_{\mathbf{i}}^*$. Following the same steps as proving Lemma 2, we can get

$$\begin{aligned}
\mathbb{E}(I_1)&=-\mathbb{E}\left(\frac{1}{h_1h_2^d}\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}+S(\mathbf{X_i})}{h_1}\right)\left(\frac{\alpha_{\mathbf{n}}\boldsymbol{\mu}^T\mathbf{X}_{\mathbf{i}}^*}{h_1}\right)K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\right)\\
&=-\mathbb{E}\Bigg(\frac{1}{h_1h_2^d}\left(\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)+\phi^{(2)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)S(\mathbf{X_i})+o(S(\mathbf{X_i}))\right)
\end{aligned}$$

$$\left(\frac{\alpha_{\mathbf{n}}\boldsymbol{\mu}^T\mathbf{X}_{\mathbf{i}}^*}{h_1}\right)K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\Bigg) = O_p(c\alpha_{\mathbf{n}}(h_1^2+h_2^2)).$$

$$\begin{aligned}\mathrm{Var}(I_1) =& \frac{1}{\tilde{\mathbf{n}}}\mathrm{Var}\left(\frac{1}{h_1h_2^d}\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}+S(\mathbf{X_i})}{h_1}\right)\left(\frac{\alpha_{\mathbf{n}}\boldsymbol{\mu}^T\mathbf{X}_{\mathbf{i}}^*}{h_1}\right)K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\right)\\ =& O_p(c^2\alpha_{\mathbf{n}}^2(\tilde{\mathbf{n}}h_1^3h_2^d)^{-1}).\end{aligned}$$

Consequently, $I_1 = O_p(c\alpha_{\mathbf{n}}(h_1^2+h_2^2)) + c\alpha_{\mathbf{n}}O_p((\tilde{\mathbf{n}}h_1^3h_2^d)^{-1/2}) = O_p(c\alpha_{\mathbf{n}}^2)$.

Similarly, we have

$$\begin{aligned}\mathbb{E}(I_2) =& \frac{1}{2}\mathbb{E}\left(\frac{1}{h_1h_2^d}\phi^{(2)}\left(\frac{\varepsilon_{\mathbf{i}}+S(\mathbf{X_i})}{h_1}\right)\left(\frac{\alpha_{\mathbf{n}}\boldsymbol{\mu}^T\mathbf{X}_{\mathbf{i}}^*}{h_1}\right)^2K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\right)\\ =& \frac{1}{2}\mathbb{E}\Bigg(\frac{1}{h_1h_2^d}\left(\phi^{(2)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)+\phi^{(3)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)S(\mathbf{X_i})+o(S(\mathbf{X_i}))\right)\\ & \left(\frac{\alpha_{\mathbf{n}}\boldsymbol{\mu}^T\mathbf{X}_{\mathbf{i}}^*}{h_1}\right)^2K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\Bigg) = O_p(c^2\alpha_{\mathbf{n}}^2).\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(I_2) =& \frac{1}{\tilde{\mathbf{n}}}\mathrm{Var}\left(\frac{1}{2h_1h_2^d}\phi^{(2)}\left(\frac{\varepsilon_{\mathbf{i}}+S(\mathbf{X_i})}{h_1}\right)\left(\frac{\alpha_{\mathbf{n}}\boldsymbol{\mu}^T\mathbf{X}_{\mathbf{i}}^*}{h_1}\right)^2K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\right)\\ =& O_p(c^4\alpha_{\mathbf{n}}^4(\tilde{\mathbf{n}}h_1^5h_2^d)^{-1}).\end{aligned}$$

As a result, $I_2 = O_p(c^2\alpha_{\mathbf{n}}^2) + c^2\alpha_{\mathbf{n}}^2O_p((\tilde{\mathbf{n}}h_1^5h_2^d)^{-1/2}) = O_p(c^2\alpha_{\mathbf{n}}^2)$. In addition, we get

$$\mathbb{E}(I_3) = -\frac{1}{6}\mathbb{E}\left(\frac{1}{h_1h_2^d}\phi^{(3)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)\left(\frac{\alpha_{\mathbf{n}}\boldsymbol{\mu}^T\mathbf{X}_{\mathbf{i}}^*}{h_1}\right)^3K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\right) = O_p(c^3\alpha_{\mathbf{n}}^3).$$

Based on these, we can choose c large enough such that I_2 dominates both I_1 and I_3 with probability $1-\eta$. Because the second term is negative, $\mathrm{P}\left\{\sup_{\|\boldsymbol{\mu}\|=c}Q_{\tilde{\mathbf{n}}}(\boldsymbol{\theta}+\alpha_{\mathbf{n}}\boldsymbol{\mu}) < Q_{\tilde{\mathbf{n}}}(\boldsymbol{\theta})\right\} \geqslant 1-\eta$ holds.

□

S7-2 Proof of Lemma 2.2

Define $R(\mathbf{X_i}) = S(\mathbf{X_i}) - (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathbf{X_i}$. Then, $Y_\mathbf{i} - a - \mathbf{b}^T(\mathbf{X_i} - \mathbf{x}) = \varepsilon_\mathbf{i} + R(\mathbf{X_i})$. The solution $\hat{\boldsymbol{\theta}}$ satisfies the equation

$$\begin{aligned}
&\sum_{\mathbf{i}\in I_\mathbf{n}} K_\mathbf{i} \phi_{h_1}^{(1)} \left(\varepsilon_\mathbf{i} + R(\mathbf{X_i})\right) \mathbf{X_i^*} \\
=&\sum_{\mathbf{i}\in I_\mathbf{n}} K_\mathbf{i} \left\{ \phi_{h_1}^{(1)} \left(\varepsilon_\mathbf{i}\right) + \phi_{h_1}^{(2)} \left(\varepsilon_\mathbf{i}\right) R(\mathbf{X_i}) + \frac{1}{2} \phi_{h_1}^{(3)} \left(\varepsilon_\mathbf{i}^*\right) R^2(\mathbf{X_i}) \right\} \mathbf{X_i^*} = 0,
\end{aligned}$$

where $\varepsilon_\mathbf{i}^*$ is between $\varepsilon_\mathbf{i}$ and $\varepsilon_\mathbf{i} + R(\mathbf{X_i})$. Based on the proof of Lemma 2, regarding the second term on the left-hand side of the above equation, we can obtain

$$\sum_{\mathbf{i}\in I_\mathbf{n}} K_\mathbf{i} \phi_{h_1}^{(2)} \left(\varepsilon_\mathbf{i}\right) \mathbf{X_i^*} S(\mathbf{X_i}) = \tilde{\mathbf{n}} \mathbb{E}(g^{(2)}(0 \mid \mathbf{X} = \mathbf{x})) \frac{h_2^2}{2} \Lambda + o_p(\tilde{\mathbf{n}} h_2^2),$$

$$-\sum_{\mathbf{i}\in I_\mathbf{n}} K_\mathbf{i} \phi_{h_1}^{(2)} \left(\varepsilon_\mathbf{i}\right) \mathbf{X_i^*} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathbf{X_i^*} = -\tilde{\mathbf{n}} \mathbb{E}(g^{(2)}(0|\mathbf{X} = \mathbf{x})) f(\mathbf{x}) S(1 + o_p(1)) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

where

$$\Lambda = \begin{pmatrix} f(\mathbf{x}) \sum_{i=1}^d \sum_{j=1}^d m_{ij}(\mathbf{x}) \int_{\mathbb{R}^d} u_i u_j K(\mathbf{u}) d\mathbf{u} \\ f(\mathbf{x}) \sum_{i=1}^d \sum_{j=1}^d m_{ij}(\mathbf{x}) \int_{\mathbb{R}^d} u_i u_j \mathbf{u} K(\mathbf{u}) d\mathbf{u} \end{pmatrix},$$

$$S = \begin{pmatrix} f(\mathbf{x}) \int_{\mathbb{R}^d} K(\mathbf{u}) d\mathbf{u} & f(\mathbf{x}) \int_{\mathbb{R}^d} \mathbf{u}^T K(\mathbf{u}) d\mathbf{u} \\ f(\mathbf{x}) \int_{\mathbb{R}^d} \mathbf{u} K(\mathbf{u}) d\mathbf{u} & f(\mathbf{x}) \int_{\mathbb{R}^d} \mathbf{u}\mathbf{u}^T K(\mathbf{u}) d\mathbf{u} \end{pmatrix}.$$

Also, from Theorem 2.1, we know that $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p((n h_2^d h_1^3)^{-1/2} + h_1^2 + h_2^2)$. Thus, we can achieve

$$\sup_{\mathbf{i}: \frac{\mathbf{X_i} - \mathbf{x}}{h_2} \le 1} |R(\mathbf{X_i})| \le \sup_{\mathbf{i}: \frac{\mathbf{X_i} - \mathbf{x}}{h_2} \le 1} [|S(\mathbf{X_i})| + |(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathbf{X_i}|] = O_p(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|) = o_p(1).$$

$$\sup_{\mathbf{i}:\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\leq 1} |R(\mathbf{X_i})|^2 = o_p(1)O_p(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|) = o_p(\alpha_{\mathbf{n}}).$$

Meanwhile, similar to the proof of Lemma 2, for $j = 0$ or 1, we have

$$\mathbb{E}\left[K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)^j\right] = \int_{\mathbb{R}^d}\frac{1}{h_2^d}K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)^j f(\mathbf{x})d\mathbf{x}$$

$$= f(\mathbf{x})\int_{\mathbb{R}^d}\mathbf{u}^j K(\mathbf{u})d\mathbf{u} + o(1).$$

Hence, in terms of the third term $\sum_{\mathbf{i}\in I_{\mathbf{n}}} K_{\mathbf{i}}\phi_{h_1}^{(3)}(\varepsilon_{\mathbf{i}}^*)R^2(\mathbf{X_i})\mathbf{X_i^*}$, we get

$$\mathbb{E}\left(\sum_{\mathbf{i}\in I_{\mathbf{n}}} K_{\mathbf{i}}\phi_{h_1}^{(3)}(\varepsilon_{\mathbf{i}}^*)R^2(\mathbf{X_i})\mathbf{X_i^*}\right) = o(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|)\sum_{\mathbf{i}\in I_{\mathbf{n}}}\mathbb{E}\left(K_{\mathbf{i}}\phi_{h_1}^{(3)}(\varepsilon_{\mathbf{i}}^*)\mathbf{X_i^*}\right) = o_p(\alpha_{\mathbf{n}}).$$

$$\mathrm{Var}\Big(\sum_{\mathbf{i}\in I_{\mathbf{n}}} K_{\mathbf{i}}\phi_{h_1}^{(3)}(\varepsilon_{\mathbf{i}}^*)R^2(\mathbf{X_i})\mathbf{X_i^*}\Big) = o(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|^2)\mathrm{Var}\Big(\sum_{\mathbf{i}\in I_{\mathbf{n}}}\left(K_{\mathbf{i}}\phi_{h_1}^{(3)}(\varepsilon_{\mathbf{i}}^*)\mathbf{X_i^*}\right)\Big)$$

$$= o_p(\alpha_{\mathbf{n}}^2).$$

Thus, the second term will dominate the third term. With the definition of $W_{\tilde{\mathbf{n}}}$, we obtain

$$\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0 = \frac{h_2^2}{2}S^{-1}\Lambda(1+o_p(1)) + \frac{S^{-1}W_{\tilde{\mathbf{n}}}}{\tilde{\mathbf{n}}g^{(2)}(0\mid\mathbf{X}=\mathbf{x})}(1+o_p(1)).$$

□

S7-3 Proof of Theorem 2.2

The main idea in the proof is to utilize the block decomposition method. Based on Lemma 2, we know

$$\frac{1}{h_1^2h_2^d}\sum_{\mathbf{i}\in I_{\mathbf{n}}}\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)$$

$$= -\tilde{\mathbf{n}}\frac{h_1^2}{2}f(\mathbf{x})g^{(3)}(0\mid\mathbf{X}=\mathbf{x})\int_{\mathbb{R}^d}K(\mathbf{u})d\mathbf{u} + o_p(1).$$

Also, by directly calculating, we can have

$$\begin{aligned}&\frac{1}{h_1^2 h_2^d}\sum_{\mathbf{i}\in I_{\mathbf{n}}}\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\\ =&-\tilde{\mathbf{n}}\frac{h_1^2}{2}f(\mathbf{x})g^{(3)}(0\mid\mathbf{X}=\mathbf{x})\int_{\mathbb{R}^d}\mathbf{u}K(\mathbf{u})d\mathbf{u}+o_p(1).\end{aligned}$$

Then, we can obtain

$$\mathbb{E}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)=\left(\frac{h_2^2}{2}S^{-1}\Lambda-\frac{h_1^2}{2}S^{-1}\frac{g^{(3)}(0\mid\mathbf{X}=\mathbf{x})}{g^{(2)}(0\mid\mathbf{X}=\mathbf{x})}\Gamma\right)(1+o_p(1)),$$

where $\Gamma=\begin{pmatrix} f(\mathbf{x})\int_{\mathbb{R}^d}K(\mathbf{u})d\mathbf{u}\\ f(\mathbf{x})\int_{\mathbb{R}^d}\mathbf{u}K(\mathbf{u})d\mathbf{u}\end{pmatrix}$.

We consider the variance of $\frac{1}{\tilde{\mathbf{n}}h_1^2h_2^d}\sum_{\mathbf{i}\in I_{\mathbf{n}}}\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\mathbf{X_i^*}$. By calculating, we have

$$\begin{aligned}&\mathrm{Var}\left(\frac{1}{\tilde{\mathbf{n}}h_1^2h_2^d}\sum_{\mathbf{i}\in I_{\mathbf{n}}}\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\mathbf{X_i^*}\right)\\ =&\frac{1}{\tilde{\mathbf{n}}h_1^4h_2^{2d}}\iint\phi^2\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)^2K^2\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\mathbf{X_i^*}\mathbf{X_i^{*T}}g(\varepsilon\mid\mathbf{X})d\varepsilon dF(\mathbf{X})\\ &(1+o_p(1))=\frac{\int\phi^2(t)t^2dt}{\tilde{\mathbf{n}}h_1^3h_2^d}g(0\mid\mathbf{X}=\mathbf{x})f(\mathbf{x})\Sigma(1+o_p(1)),\end{aligned}$$

where $\Sigma=\begin{pmatrix}\int_{\mathbb{R}^d}K^2(\mathbf{u})d\mathbf{u} & \int_{\mathbb{R}^d}\mathbf{u}^TK^2(\mathbf{u})d\mathbf{u}\\ \int_{\mathbb{R}^d}\mathbf{u}K^2(\mathbf{u})d\mathbf{u} & \int_{\mathbb{R}^d}\mathbf{u}\mathbf{u}^TK^2(\mathbf{u})d\mathbf{u}\end{pmatrix}$. Then, we get

$$\mathrm{Var}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)=\frac{\int\phi^2(t)t^2dt}{\tilde{\mathbf{n}}h_1^3h_2^d}\frac{g(0\mid\mathbf{X}=\mathbf{x})}{(g^{(2)}(0\mid\mathbf{X}=\mathbf{x}))^2}f(\mathbf{x})S^{-1}\Sigma S^{-1}(1+o_p(1)).$$

To prove Theorem 2.2, it is sufficient to show that

$$\sqrt{h_1^3 h_2^d \tilde{\mathbf{n}}}\frac{1}{\tilde{\mathbf{n}} h_1^2 h_2^d}\sum_{\mathbf{i}\in I_{\mathbf{n}}}\left[\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\mathbf{X_i^*}-\mathbb{E}(\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)\right.$$

$$\left.K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\mathbf{X_i^*}\right)\Big] \xrightarrow{d} \mathcal{N}\Big(0, \int \phi^2(t)t^2dt\mathbb{E}(g(0\mid \mathbf{X}=\mathbf{x}))f(\mathbf{x})\Sigma\Big).$$

Using Slutsky's theorem, Theorem 2.2 follows directly. After that, for any unit vector $C \in \mathbb{R}^{d+1}$, we need to prove the following asymptotic result

$$\{C^T \mathrm{Cov}(W_{\tilde{\mathbf{n}}}^*)C\}^{-1/2}\{C^T W_{\tilde{\mathbf{n}}}^* - C^T\mathbb{E}(W_{\tilde{\mathbf{n}}}^*)\} \xrightarrow{d} \mathcal{N}(0,1),$$

where $W_{\tilde{\mathbf{n}}}^* - \mathbb{E}(W_{\tilde{\mathbf{n}}}^*) = \sqrt{\frac{1}{\tilde{\mathbf{n}}h_1h_2^d}}\sum_{\mathbf{i}\in I_{\mathbf{n}}}\Big[\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\mathbf{X_i^*} - \mathbb{E}\Big(\phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right)$ $K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right)\mathbf{X_i^*}\Big)\Big] = \sum_{\mathbf{i}\in I_{\mathbf{n}}}\Delta_{\mathbf{i}}$. Now, we decompose $C^T W_{\tilde{\mathbf{n}}}^* - \mathbb{E}(C^T W_{\tilde{\mathbf{n}}}^*)$ into smaller pieces involving "large" and "small" blocks. More specifically, we consider

$$U(1,\mathbf{n},\mathbf{j}) := \sum_{i_k=j_k(p_k+q)+1,k=1,\cdots,N}^{j_k(p_k+q)+p_k}\Delta_{\mathbf{i}},$$

$$U(2,\mathbf{n},\mathbf{j}) := \sum_{i_k=j_k(p_k+q)+1,k=1,\cdots,N-1}^{j_k(p_k+q)+p_k}\sum_{i_N=j_N(p_N+q)+p_N+1}^{(j_N+1)(p_N+q)}\Delta_{\mathbf{i}},$$

$$U(3,\mathbf{n},\mathbf{j}) := \sum_{i_k=j_k(p_k+q)+1,k=1,\cdots,N-2}^{j_k(p_k+q)+p_k}\sum_{i_{N-1}=j_{N-1}(p_{N-1}+q)+p_{N-1}+1}^{(j_{N-1}+1)(p_{N-1}+q)}\sum_{i_N=j_N(p_N+q)+1}^{j_N(p_N+q)+p_N}\Delta_{\mathbf{i}},$$

$$U(4,\mathbf{n},\mathbf{j}) := \sum_{i_k=j_k(p_k+q)+1,k=1,\cdots,N-2}^{j_k(p_k+q)+p_k}\sum_{i_{N-1}=j_{N-1}(p_{N-1}+q)+p_{N-1}+1}^{(j_{N-1}+1)(p_{N-1}+q)}\sum_{i_N=j_N(p_N+q)+p_N+1}^{(j_N+1)(p_N+q)}\Delta_{\mathbf{i}},$$

and so on. It is noticed that

$$U\left(2^N-1,\mathbf{n},\mathbf{j}\right) := \sum_{i_k=j_k(p_k+q)+p_k+1,k=1,\cdots,N-1}^{(j_k+1)(p_k+q)}\sum_{i_N=j_N(p_N+q)+1}^{j_N(p_N+q)+p_N}\Delta_{\mathbf{i}},$$

$$U\left(2^N, \mathbf{n}, \mathbf{j}\right) := \sum_{i_k = j_k(p_k+q)+p_k+1, k=1,\cdots,N}^{(j_k+1)(p_k+q)} \Delta_{\mathbf{i}}.$$

Without loss of generality, assume that, for some integers $r_1, \cdots, r_N$, $\mathbf{n} = (n_1, \cdots, n_N)$ is such that $n_1 = r_1 (p_1 + q), \cdots, n_N = r_N (p_N + q)$, with $r_k \to \infty$ for all $k = 1, \cdots, N$. For each integer $1 \leq i \leq 2^N$, define

$$T(\mathbf{n}, i) := \sum_{j_k = 0, k=1,\cdots,N}^{r_k - 1} U(i, \mathbf{n}, \mathbf{j}).$$

Then, we have

$$C^T W^*_{\tilde{\mathbf{n}}} = \sum_{i=1}^{2^N} T(\mathbf{n}, \mathbf{x}, i).$$

Note that $T(\mathbf{n}, 1)$ is the sum of the random variables $\Delta_{\mathbf{i}}$ over "large" blocks, whereas $T(\mathbf{n}, \mathbf{x}, i), 2 \leq i \leq 2^N$, are sums over "small" blocks. If it is not the case that $n_1 = r_1 (p_1 + q), \cdots, n_N = r_N (p_N + q)$ for some integers $r_1, \cdots, r_N$, then an additional term $T\left(\mathbf{n}, 2^N + 1\right)$ containing all the $\Delta_{\mathbf{i}}$'s that are not included in the big or small blocks can be considered. This term will not change the proof much. As $\mathbf{n} \to \infty$, we obtain

$$Q_1 := \left| \mathbb{E}[\exp[iuT(\mathbf{n}, 1)]] - \prod_{j_k = 0, k=1,\cdots,N}^{r_k - 1} \mathbb{E}[\exp[iuU(1, \mathbf{n}, \mathbf{j})]] \right| \to 0,$$

$$Q_2 := \mathbb{E}\left(\sum_{i=2}^{2^N} T(\mathbf{n}, i) \right)^2 \to 0,$$

$$Q_3 := \sum_{j_k = 0, k=1,\cdots,N}^{r_k - 1} \mathbb{E}[U(1, \mathbf{n}, \mathbf{j})]^2 \to \int \phi^2(t) t^2 dt \mathbb{E}(g(0 \mid \mathbf{X} = \mathbf{x})) f(\mathbf{x}) C^T \Sigma C,$$

$$Q_4 := \sum_{j_k=0,k=1,\cdots,N}^{r_k-1} \mathbb{E}\Big[(U(1,\mathbf{n},\mathbf{j}))^2 I\Big\{|U(1,\mathbf{n},\mathbf{j})| > \epsilon \int \phi^2(t)t^2dt$$

$$\mathbb{E}(g(0 \mid \mathbf{X}=\mathbf{x}))f(\mathbf{x})C^T\Sigma C\Big\}\Big] \to 0,$$

for every $\epsilon > 0$. We then prove equations Q_1-Q_4.

Rank the random variables $U(1,\mathbf{n},\mathbf{x},\mathbf{j})$ in an arbitrary manner and refer to them as $\widetilde{U}_1,\cdots,\widetilde{U}_M$. Note that $M = \prod_{k=1}^N r_k = \hat{\mathbf{n}}\left\{\prod_{k=1}^N (p_k+q)\right\}^{-1} \leq \hat{\mathbf{n}}/p$, where $p = \prod_{k=1}^N p_k$. Let

$$\ell(1,\mathbf{n},\mathbf{x},\mathbf{j}) := \{\mathbf{i} : j_k\,(p_k+q)+1 \leq i_k \leq j_k\,(p_k+q)+p_k, k=1,\cdots,N\}\,.$$

Following the results in Lemma 4, we have

$$\begin{aligned}
Q_1 \leqslant & \sum_{k=1}^{M-1}\sum_{j=k+1}^{M} \left|\mathbb{E}\left(\exp\left[iu\tilde{U}_k\right]-1\right)\left(\exp\left[iu\tilde{U}_j\right]-1\right)\prod_{s=j+1}^{M}\exp\left[iu\tilde{U}_s\right]\right. \\
& \left. -\,\mathbb{E}\left(\exp\left[iu\tilde{U}_k\right]-1\right)\mathbb{E}\left(\exp\left[iu\tilde{U}_j\right]-1\right)\prod_{s=j+1}^{M}\exp\left[iu\tilde{U}_s\right]\right| \\
\leq & \sum_{k=1}^{M-1}\sum_{j=k+1}^{M} \left|\mathbb{E}\left(\exp\left[iu\tilde{U}_k\right]-1\right)\left(\exp\left[iu\tilde{U}_j\right]-1\right)\right. \\
& \left. -\,\mathbb{E}\left(\exp\left[iu\tilde{U}_k\right]-1\right)\mathbb{E}\left(\exp\left[iu\tilde{U}_j\right]-1\right)\right| \\
\leq & C\sum_{k=1}^{M-1}\min\{p,(M-k)p\}\varphi(q) \leq CMp\varphi(q) \leq C\tilde{\mathbf{n}}\varphi(q) \to 0.
\end{aligned}$$

In order to prove Q_2, it is enough to show that

$$\mathbb{E}[T^2(\mathbf{n},i)] \to 0 \text{ for any } 2 \leq i \leq 2^N.$$

Without loss of generality, we consider $\mathbb{E}\,[T^2(\mathbf{n},2)]$. Ranking the random variables $U(2,\mathbf{n},\mathbf{j})$ in an arbitrary manner and referring them as

$\widehat{U}_1, \cdots, \widehat{U}_M$, we can get

$$\mathbb{E}\left[T^2(\mathbf{n},2)\right] = \sum_{i=1}^{M} \mathrm{Var}\left(\widehat{U}_i\right) + 2 \sum_{1\le i<j\le M} \mathrm{Cov}\left(\widehat{U}_i, \widehat{U}_j\right) = B_1 + B_2.$$

We shall prove $B_1 \to 0$ and $B_2 \to 0$ below.

As we know $\Delta_{\mathbf{i}} \approx (\tilde{\mathbf{n}} h_1 h_2^d)^{-1/2} \phi^{(1)}\left(\frac{\varepsilon_{\mathbf{i}}}{h_1}\right) K\left(\frac{\mathbf{X_i}-\mathbf{x}}{h_2}\right) \mathbf{X}_{\mathbf{i}}^*$, then

$$\mathrm{Var}(\Delta_{\mathbf{i}}) \approx (\tilde{\mathbf{n}} h_1 h_2^d)^{-1} (h_1 h_2^d) \iint \phi^2(t) t^2 K^2(\mathbf{u}) \mathbf{X}_{\mathbf{i}}^* \mathbf{X}_{\mathbf{i}}^{*T} g(t h_1 \mid \mathbf{X}=\mathbf{x})$$

$$f(h_2\mathbf{u}+\mathbf{x}) dt d\mathbf{u} (1+o_p(1)) \le C\tilde{\mathbf{n}}^{-1}.$$

According to the calculation, we obtain

$$\mathbb{E}(\Delta_{\mathbf{i}}\Delta_{\mathbf{j}}) \approx (\tilde{\mathbf{n}} h_1 h_2^d)^{-1} \iiint \phi(t) t \phi(s) s K(\mathbf{u}) K(\mathbf{v}) \mathbf{X}_{\mathbf{i}}^* \mathbf{X}_{\mathbf{j}}^{*T} g(t h_1, s h_1 \mid \mathbf{X}=\mathbf{x})$$

$$f(h_2\mathbf{u}+\mathbf{x}, h_2\mathbf{v}+\mathbf{x}) dt ds d\mathbf{u} d\mathbf{v} (1+o_p(1)) \le C\left(\tilde{\mathbf{n}} h_1 h_2^d\right).$$

Based on the results from Lemma 3, we have

$$B_1 = \mathbb{E}\left[\left(\sum_{i_k=1,k=1,\cdots,N-1}^{p_k} \sum_{i_N=1}^{q} \Delta_{\mathbf{i}}\right)^2\right] + \sum_{\mathbf{i}\neq\mathbf{j}\in\mathfrak{F}} \mathbb{E}\left[\Delta_{\mathbf{i}}\Delta_{\mathbf{j}}\right] \le C\Bigg(\left(\prod_{k=1}^{N-1} p_k\right) q$$

$$\mathrm{Var}\left\{\Delta_{\mathbf{i}}\right\} + h_1^4 h_2^d \tilde{\mathbf{n}}^{-1} \left(\prod_{k=1}^{N-1} p_k c_{\mathbf{n}k}\right) q c_{\mathbf{n}N} + \tilde{\mathbf{n}} (h_1 h_2^d)^{-1} (h_1 h_2^d)^{2/(2+\delta)} \left(\prod_{k=1}^{N-1} p_k\right)$$

$$q \sum_{k=1}^{N} \sum_{t=c_{\mathbf{n}k}}^{\|\mathbf{n}\|} t^{N-1} \{\varphi(t)\}^{\delta/(2+\delta)}\Bigg) \le C\tilde{\mathbf{n}}^{-1} \left(\prod_{k=1}^{N-1} p_k\right) q \le C(q/p_N) \to 0,$$

where $\mathfrak{F} = \mathfrak{F}(\mathbf{p},q) := \{\mathbf{i},\mathbf{j} : 1 \le i_k, j_k \le p_k, k = 1,\cdots,N-1, \text{ and } 1 \le i_N, j_N \le q\}$.

Set $\ell(2,n,\mathbf{x},\mathbf{j}) := \{\mathbf{i} : j_k(p_k+q)+1 \le i_k \le j_k(p_k+q)+p_k, 1 \le k \le N-1, j_N(p_N+q)+p_N+1 \le i_N \le (j_N+1)(p_N+q)\}$. Then, $U(2,\mathbf{n},\mathbf{j}) =$

$\sum_{\mathbf{i}\in\ell(2,\mathbf{n},\mathbf{i})}\Delta_{\mathbf{i}}$. Since $p_k > q$, if $\mathbf{i}$ and $\mathbf{i}'$ belong to two distinct sets $\ell(2,\mathbf{n},\mathbf{x},\mathbf{j})$ and $\ell\left(2,\mathbf{n},\mathbf{x},\mathbf{j}'\right)$, we have $\|\mathbf{i}-\mathbf{i}'\| > q$. In view of Lemma 3 by taking $n_k = p_k$ and $n_N = q$, following the same arguments as proving B_1, we get

$$
\begin{aligned}
|B_2| &\leq 2\sum_{1\leq i<j\leq M}|\mathrm{Cov}\left(\widehat{U}_i,\widehat{U}_j\right)| \leq C\sum_{\{\mathbf{i},\mathbf{j}:\|\mathbf{i}-\mathbf{j}\|\geq q,1\leq i_k,j_k\leq n_k\}}|\mathbb{E}\left[\Delta_{\mathbf{i}}\Delta_{\mathbf{j}}\right]| \\
&\leq C(h_1h_2^d)^{-1}\sum_{\{\mathbf{i},\mathbf{j}:\|\mathbf{i}-\mathbf{j}\|\geq q,1\leq i_k,j_k\leq n_k\}}(h_1h_2^d)^{2/(2+\delta)}\{\varphi(\|\mathbf{j}-\mathbf{i}\|)\}^{\delta/(2+\delta)} \\
&\leq C(h_1h_2^d)^{-\delta/(2+\delta)}\left(\prod_{k=1}^{N}n_k\right)\left(\sum_{t=q}^{\|\mathbf{n}\|}t^{N-1}\{\varphi(t)\}^{\delta/(2+\delta)}\right).
\end{aligned}
$$

Take $c_{\mathbf{n}k}^a = (h_1h_2^d)^{-\delta/(2+\delta)} \to \infty$. As $q(h_1h_2^d)^{\delta/a(2+\delta)} > 1$, so that $c_{\mathbf{n}k} < q \leq p_k$. Thus, Q_2 follows.

$\mathbf{Q_3}$: The result can be easily obtained by following the initial procedures in this proof and the arguments for Q_2. Particularly, let $S'_{\mathbf{n}} := T(\mathbf{n},1)$ and $S''_{\mathbf{n}} := \sum_{i=2}^{2^N}T(\mathbf{n},i)$. Then, $S'_{\mathbf{n}}$ is a sum of $Y_{\mathbf{j}}$'s over the "large" blocks and $S''_{\mathbf{n}}$ over the "small" ones. We can write

$$
\mathbb{E}\left(S'_{\mathbf{n}}\right)^2 = \mathbb{E}S_{\mathbf{n}}^2 + \mathbb{E}\left(S''_{\mathbf{n}}\right)^2 - 2\mathbb{E}S_{\mathbf{n}}S''_{\mathbf{n}}.
$$

In the beginning, we have shown that $\mathbb{E}\left[|S_{\mathbf{n}}|^2\right] \to \int\phi^2(t)t^2dt\mathbb{E}(g(0 \mid \mathbf{X}=\mathbf{x}))f(\mathbf{x})\Sigma$. Also, Q_2 implies $\mathbb{E}\left(S''_{\mathbf{n}}\right)^2 \to 0$. To demonstrate

$$
\mathbb{E}\left[|S'_{\mathbf{n}}|^2\right] \to \int\phi^2(t)t^2dt\mathbb{E}(g(0 \mid \mathbf{X}=\mathbf{x}))f(\mathbf{x})\Sigma,
$$

it is sufficient to show that $\tilde{\mathbf{n}}^{-1}\mathbb{E}S_{\mathbf{n}}S''_{\mathbf{n}} \to 0$, as by Cauchy-Schwartz's inequality, we can write

$$|\mathbb{E}S_{\mathbf{n}}S''_{\mathbf{n}}| \leq \mathbb{E}|S_{\mathbf{n}}S''_{\mathbf{n}}| \leq \left(\mathbb{E}S^2_{\mathbf{n}}\right)^{1/2}\left(\mathbb{E}S''^2_{\mathbf{n}}\right)^{1/2}.$$

Recall that $T(\mathbf{n},1)=\sum_{\mathbf{j}\in\mathcal{J}}U(1,\mathbf{n},\mathbf{j})$. We then have

$$\begin{aligned}\mathbb{E}\left[|S'_{\mathbf{n}}|^2\right] = & \sum_{j_k=0,k=1,\cdots,N}^{r_k-1}\mathbb{E}\left[U^2(1,\mathbf{n},\mathbf{j})\right]\\ & +\sum_{\mathbf{i}\neq\mathbf{j}\in\mathcal{J}^*}\mathrm{Cov}(U(1,\mathbf{n},\mathbf{j}),U(1,\mathbf{n},\mathbf{i})),\end{aligned}$$

where $\mathcal{J}^*=\mathcal{J}^*(\mathbf{p},q):=\{\mathbf{i},\mathbf{j}:1\leq i_k,j_k\leq r_k-1,k=1,\cdots,N\}$. By the same argument used in showing B_2, we can prove that

$$\begin{aligned}&C(h_1h_2^d)^{-\delta/(2+\delta)}\sum_{\|\mathbf{i}\|>q}\sum_{i_k=1,k=1,\cdots,N}^{n_k-1}\{\varphi(\|\mathbf{i}\|)\}^{\delta/(2+\delta)}\\ \leq & C(h_1h_2^d)^{-\delta/(2+\delta)}\left(\sum_{t=q}^{\infty}t^{N-1}\{\varphi(t)\}^{\delta/(2+\delta)}\right)\to 0.\end{aligned}$$

$\mathbf{Q_4}$: At first, we know $|U(1,\mathbf{n},\mathbf{j})|\leq Cp(\tilde{\mathbf{n}}h_1h_2^d)^{-1/2}$. It follows that

$$\begin{aligned}Q_4\leq & Cp^2(\tilde{\mathbf{n}}h_1h_2^d)^{-1}\sum_{j_k=0,k=1,\cdots,N}^{r_k-1}P\Big[|U(1,\mathbf{n},\mathbf{j})|\\ & >\epsilon\left(\int\phi^2(t)t^2dt\mathbb{E}(g(0\mid\mathbf{X}=\mathbf{x}))f(\mathbf{x})C^T\Sigma C\right)^{1/2}\tilde{\mathbf{n}}^{1/2}\Big].\end{aligned}$$

Moreover,

$$\begin{aligned}&|U(1,\mathbf{n},\mathbf{j})|\Big/\left(\left(\int\phi^2(t)t^2dt\mathbb{E}(g(0\mid\mathbf{X}=\mathbf{x}))f(\mathbf{x})C^T\Sigma C\right)^{1/2}\tilde{\mathbf{n}}^{1/2}\right)\\ \leq & Cp(\tilde{\mathbf{n}}h_1h_2^d)^{-1/2}\to 0,\end{aligned}$$

since $p=[(\tilde{\mathbf{n}}h_1h_2^d)^{1/2}/s_{\mathbf{n}}]$, where $s_{\mathbf{n}}\to\infty$. Thus, we can obtain

$$P\left[|U(1,\mathbf{n},\mathbf{j})|>\epsilon\left(\int\phi^2(t)t^2dt\mathbb{E}(g(0\mid\mathbf{X}=\mathbf{x}))f(\mathbf{x})C^T\Sigma C\right)^{1/2}\tilde{\mathbf{n}}^{1/2}\right]=0$$

for sufficiently larger $\mathbf{n}$ at all $\mathbf{j}$.

□

Bibliography

Feng, Y., Fan, J., and Suykens, J. A. K. (2020). A Statistical Learning Approach to Modal Regression. *Journal of Machine Learning Research*, 21 (2), 1-35.

Firpo, S., Fortin, N. M., and Lemieux, T. (2009). Unconditional Quantile Regressions. *Econometrica*, 77 (3), 953-973.

Hallin, M., Lu, Z., and Tran, L. T. (2004). Local Linear Spatial Regression. *The Annals of Statistics*, 32 (6), 2469-2500.

Hallin, M., Lu, Z., and Yu, K. (2009). Local Linear Spatial Quantile Regression. *Bernoulli*, 15 (3), 659-686.

Lu, Z., Steinskog, D. J., Tjøstheim, D., and Yao, Q. (2009). Adaptively Varying Coefficient Spatiotemporal Models. *Journal of the Royal Statistical Society Series B*, 71 (4), 859-880.

Ortega, J. M. and Rheinboldt, W. C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press.

Su, L. and Yang, Z. (2011). Instrumental Variable Quantile Estimation of Spatial Autoregressive Model. *Research Collection School of Economics Working Paper May 2011*.

Tran, L. T. (1990). Kernel Density Estimation on Random Fields. *Journal of Multivariate Analysis*, 34 (1), 37-53.

Ullah, A., Wang, T., and Yao, W. (2022). Nonlinear Modal Regression for Dependent Data with Application for Predicting COVID-19. *Journal of the Royal Statistical Society Series A*, 185 (3), 1424-1453.

Ullah, A., Wang, T., and Yao, W. (2023). Semiparametric Partially Linear Varying Coefficient Modal Regression. *Journal of Econometrics*, 235 (2), 1001-1026.

Yu, K. and Lu, Z. (2004). Local Linear Additive Quantile Regression. *Scandinavian Journal of Statistics*, 31 (3), 333-346.