Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 8(6): 13923–13942.
DOI:10.3934/math.2023713
Received: 04 February 2023
Revised: 13 March 2023
Accepted: 17 March 2023
Published: 13 April 2023

Research article

Adaptive estimation for spatially varying coefficient models

Heng Liu and Xia Cui*

School of Economics and Statistics, Guangzhou University, Guangzhou 510006, China

* **Correspondence:** Email: cuixia@gzhu.edu.cn.

Abstract: In this paper, a new adaptive estimation approach is proposed for the spatially varying coefficient models with unknown error distribution, unlike geographically weighted regression (GWR) and local linear geographically weighted regression (LL), this method can adapt to different error distributions. A generalized Modal EM algorithm is presented to implement the estimation, and the asymptotic property of the estimator is established. Simulation and real data results show that the gain of the new adaptive method over the GWR and LL estimation is considerable for the error of non-Gaussian distributions.

Keywords: adaptive estimation; generalized Modal EM algorithm; geographically weighted regression; spatially varying coefficient models
Mathematics Subject Classification: 62G05

1. Introduction

In spatial data analysis, a common problem is determining the nature of the relationship between variables. In many cases, a simple global model often fails to explain the relationships between certain sets of variables, as the relationships between them may change with the change of position, which is known as spatial heterogeneity. In order to deal with this heterogeneity, the model needs to reflect the structure of spatial variation in the data. Suppose that the spatial data of n positions are randomly selected in the spatial region $D \subseteq R^2$, let $\boldsymbol{u}_i = (u_{i1}, u_{i2})^\top \in D$ is the position of the point $i, i = 1, \cdots, n$, y_i is the response variable, $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^\top$ is the explanatory variable and $x_{i1} \equiv 1$, allowing a varying intercept in the model. $\{y_i, \mathbf{x}_i, u_i\}$ satisfy the following spatially varying coefficient models (SVCM) [1–3]:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}(\boldsymbol{u}_i) + \varepsilon_i = \sum_{k=1}^{p} x_{ik}\beta_k(\boldsymbol{u}_i) + \varepsilon_i, \qquad i = 1, 2, \cdots, n, \tag{1.1}$$

where $\boldsymbol{\beta}(\boldsymbol{u}_i) = (\beta_1(\boldsymbol{u}_i), \beta_2(\boldsymbol{u}_i), \cdots, \beta_p(\boldsymbol{u}_i))^\top$ is a vector of p-dimensional unknown space-varying functional coefficients defined on D, ε_i is an independent and and identically distributed random noise, with $E(\varepsilon_i) = 0, var(\varepsilon_i) = \sigma^2$, and are independent of $\mathbf{x}_i$. Over the past few decades, SVCM has been widely used in geography [4], econometrics [5], meteorology [6], and environmental science [7]. When $\beta_k(\cdot)$ is a univariate function, the model (1.1) is a varying coefficient model and has been extensively studied [8, 9]. In this study, $\beta_k(\cdot)$ is a bivariate function of the location-specific, and our main goal is to estimate $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)$ and explore the spatial heterogeneity of regression relations based on the given observations $\{(y_i, \mathbf{x}_i, \boldsymbol{u}_i)\}_{i=1}^n$.

In the rich literature on how to estimate the regression coefficients of SVCM, the Bayesian approach and the smoothing approach are two competing methods. Firstly, the Bayesian approach is an important spatial modeling method that assumes that the regression coefficients obey a certain prior distribution and calculates their posterior distribution for estimation and inference. For example, Gelfand et al. [10] developed a Bayesian hierarchical framework of spatial point reference data by formulating a Gaussian process for spatially varying coefficients, and Assuncao [11] introduced the Bayesian space-varying coefficient model (BVCM) for areal data. Recently, Kim and Lee [12] extended BVCM to handle mixed data with point reference data and areal data. Luo et al. [13] built the Bayesian spatially clustered coefficient (BSCC) model from the spanning trees of a graph. However, Bayesian methods require careful selection of prior distributions and face the high computational cost issue. Secondly, the smoothing method is a traditional framework for regression, divided into kernel smoothing and smoothing splines. For example, Fotheringham et al. [1] adopted a locally weighted least squares method of constructing weights by spatial kernel functions, namely geographically weighted regression (GWR), which is essentially a local constant kernel smoother. Mu et al. [14] used binary splines over triangulation to estimate regression coefficients, which solves the problem of inappropriate smoothness of complex regional boundary features and processes large data sets quickly and effectively enough. Yet, the kernel-based method needs to solve an optimization problem at each sample position which is computationally intensive, and smoothing splines method inference for spatially varying coefficients relies on a bootstrap method.

Currently, there are also numerous studies on variable selection in SVCM. Shin et al. [15] proposed penalized quasi-likelihood methods with spatial dependence. Wang and Sun [16] represented the space-varying coefficients as a combination of local polynomials at anchor points and applying the least squares with an additive form of lasso and fused-lasso penalties. Li and Sang [17] proposed a spanning tree graph fused lasso-based spatially clustered coefficient regression (SCC) model with the assumption of spatial clusters, and the regularization term of the SCC model is generalized by a chain graph guided fusion penalty plus a group lasso penalty [18]. However, each of these methods estimated space-varying coefficients by the least squares criterion, corresponding to the likelihood function when the error term is normally distributed. In practice, the error density was unknown, so it is not appropriate to use the least squares method, which will lose some efficiency, but the adaptive estimation method provides an alternative way.

The adaptive estimation method was first studied to consider the problem of estimating and inferring an infinite dimensional parameter [19]. This method replaces the Gaussian density function with a nonparametric estimate of the score function of the log-likelihood estimation and proves that efficiency gain can be achieved in both varying coefficient models [20] and varying coefficient models with non-stationary covariates [21]. In this study, we propose an adaptive estimation method to estimate spatially varying coefficients, different from the least squares criterion, the logarithmic function of the new adaptive estimation method is similar to the likelihood structure of the mixed density function, without an explicit solution, and we use the generalized Modal EM (GMEM) algorithm to achieve parameter estimation [22]. Simulation results show that when the error distribution deviates from the normal distribution, the new estimation is more effective than the existing GWR estimation based on least squares. In addition, the new method is also comparable with existing GWR methods when the error is completely normal. Finally, we illustrate the effectiveness of the proposed adaptive estimation method through two real data examples.

The rest of this study is organized below. In Section 2, the adaptive estimation of spatial varying coefficient models and the generalized Modal EM algorithm are introduced. In Section 3, through simulation research, the proposed method is compared with the GWR method under five different error densities. In Section 4, the new method is applied to two real-world data examples. This article is briefly discussed in Section 5. All technical conditions and certifications are given in Section Appendix A.

2. Adaptive kernel estimation method

For any given $\boldsymbol{u}_0$, approximating the spatially varying coefficients by Taylor's expansion as

$$\beta_k(\boldsymbol{u}_i) \approx \beta_k(\boldsymbol{u}_0) + \dot{\boldsymbol{\beta}}_k(\boldsymbol{u}_0)(\boldsymbol{u}_i - \boldsymbol{u}_0) \overset{\triangle}{=} b_k + \boldsymbol{c}_k(\boldsymbol{u}_i - \boldsymbol{u}_0), \quad k = 0, \cdots, p, \tag{2.1}$$

where $\boldsymbol{u}_i$ is in a neighborhood of $\boldsymbol{u}_0$, $\dot{\boldsymbol{\beta}}_k(\boldsymbol{u}_0) = \{\partial(\beta_k(\boldsymbol{u})/\partial u_1, \partial(\beta_k(\boldsymbol{u})/\partial u_2\}_{\boldsymbol{u}=\boldsymbol{u}_0}$. Using the above approximation, we have the following objective function for estimating $(b_1, \cdots, b_p)$ and $(\boldsymbol{c}_1, \cdots, \boldsymbol{c}_p)$

$$\sum_{i=1}^{n} \left[y_i - \sum_{k=1}^{p} \{b_k + \boldsymbol{c}_k(\boldsymbol{u}_i - \boldsymbol{u}_0)\} x_{ik} \right]^2 K_h(||\boldsymbol{u}_i - \boldsymbol{u}_0||), \tag{2.2}$$

where $K_h(\cdot) = K(\cdot/h)/h^2$, $K(\cdot)$ is a kernel function, h is a bandwidth, and $||\boldsymbol{s}|| = (\boldsymbol{s}^\top \boldsymbol{s})^{\frac{1}{2}}$ for a vector $\boldsymbol{s}$. Throughout this study, a Gaussian kernel will be used for $K(\cdot)$. Due to the least squares in (2.2), the resulting estimate may lose some efficiency when the error distribution is not normal. Therefore, we develop an adaptive estimation procedure that can adapt to different error distributions.

Let $f(\varepsilon)$ be the density function of ε. If $f(\varepsilon)$ were known, it would be natural to estimate the parameters in (2.1) by maximizing the following log-likelihood function

$$\sum_{i=1}^{n} \log f \left[y_i - \sum_{k=1}^{p} \{b_k + \boldsymbol{c}_k(\boldsymbol{u}_i - \boldsymbol{u}_0)\} x_{ik} \right] K_h(||\boldsymbol{u}_i - \boldsymbol{u}_0||). \tag{2.3}$$

However, in practice, $f(\varepsilon)$ is generally unknown but can be replaced by a leave-one-out kernel density estimator

$$\tilde{f}_\varepsilon = \frac{1}{n} \sum_{j \neq i}^{n} K_{h_0}(\varepsilon_i - \tilde{\varepsilon}_j), \tag{2.4}$$

where $\tilde{\varepsilon}_j = y_j - \sum_{k=0}^{p} x_{jk} \tilde{\beta}_k(\boldsymbol{u}_j)$ is a preliminary estimation of ε_j based on initial estimator$\tilde{\beta}_k$, +and$\tilde{\beta}_k = \tilde{b}_k$, which can be estimated by local linear regression estimator (2.2). Let $\boldsymbol{\theta} = (b_1, \cdots, b_p, \boldsymbol{c}_1, \cdots, \boldsymbol{c}_p)^\top$. Then our proposed adaptive estimate for the parameter $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}) \tag{2.5}$$

where

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left(\frac{1}{n}\sum_{j\neq i} K_{h_0}[y_i - \sum_{k=1}^{p}\{b_k + \boldsymbol{c}_k(\boldsymbol{u}_i - \boldsymbol{u}_0)\}x_{ik} - \tilde{\varepsilon}_j]\right) K_h(\|\boldsymbol{u}_i - \boldsymbol{u}_0\|). \tag{2.6}$$

Since the logarithmic function of (2.6) has an internal sum, which is similar to the objective function from a random sample of mixed density, so there is no explicit solution. In the following, we use the generalized Modal EM algorithm proposed in Yao [22] to calculate the parameters.

Generalized Modal EM algorithm (GMEM): GMEM algorithm is the generalization of Modal EM (MEM) algorithm [23] that finds the mode of the mixture density and does nonparametric clustering. The MEM algorithm comprises two steps similar to the expectation and the maximization steps in EM algorithm, which aims at maximizing the likelihood function for finite mixture models when the model contains unobserved latent variables. Especially, suppose an m-component finite mixture density be

$$f(x) = \sum_{j=1}^{m} \pi_j f_j(x),$$

where π_j is the mixing proportions of mixture component j, and $f_j(x)$ is the density of component j. Given any initial value $x^{(0)}$, in the $(l+1)th$ step of the MEM algorithm solves a local maximum of the mixture by the following two steps:
1) let

$$p_j = \frac{\pi_j f_j(x^{(l)})}{f(x^{(l)})}, j = 1, \cdots, m,$$

2) update

$$x^{(l+1)} = \arg\max_{x} \sum_{j=1}^{m} p_j \log f_j(x).$$

The first step is the "Expectation" step where the probability of each mixture component $j, 1 \leq j \leq m$, at the current point $x^{(r)}$ is computed. The second step is the "Maximization" step, similar to EM algorithm, which is usually much easier than the original objective function. Detailed properties of MEM algorithms refer to Li et al. [23]. Yao [22] proves that the MEM algorithm can be applied to maximize a general mixture-type objective function

$$f(x) = \sum_{j=1}^{m} w_j \left[\log\{\sum_{k=1}^{K} a_{jk} f_{jk}(x)\}\right] \tag{2.7}$$

where w_k and a_{kl} are known positive constants, $f_{jk}(x)$ is positive known function, when $j = 1$, the objective function (2.7) is simplified to

$$f(x) = w_1 \log\{\sum_{k=1}^{K} a_{1k} f_{1k}(x)\} \propto \sum_{k=1}^{K} a_{1k} f_{1k}(x),$$

Therefore, the MEM algorithm in Eq (2.7) is a special case of the generalized Modal EM algorithm (GMEM) if $\sum_{k=1}^{K} a_{1k} = 1$ and $f_{1k}(x)$ are density functions. Specifically, given the initial value $x^{(0)}$, in

the $(l+1)th$ step of the GMEM algorithm are following:
E-step: let

$$p_{jk}^{(l+1)}=\frac{a_{jk}f_{jk}(x^{(l)})}{\sum_{k=1}^{K}a_{jk}f_{jk}(x^{(l)})},\, j=1,\cdots,m, k=1,\cdots,K,$$

M-step: update

$$x^{(l+1)}=\arg\max_{x}\sum_{j=1}^{m}\sum_{k=1}^{K}\{w_j p_{jk}^{l+1}\log f_{jk}(x)\}.$$

In this study, we note that the objective function $Q(\boldsymbol{\theta})$ of (2.6) has the mixture form of (2.7). Specially, $K_h(\boldsymbol{u}_i-\boldsymbol{u}_0), \frac{1}{n}, K_{h_0}[y_i-\sum_{k=1}^{p}\{b_k+\boldsymbol{c}_k(\boldsymbol{u}_i-\boldsymbol{u}_0)\}x_{ik}-\tilde{\varepsilon}_j]$ in (2.6) corresponds to $w_j, a_{jk}, f_{jk}(x)$ in (2.7), respectively. Therefore, GMEM could be directly applied to estimate the parameters of b_k, $\boldsymbol{c}_k$ in (2.6). Let $\boldsymbol{\theta}^{(0)}$ be the initial estimator obtained by minimizing (2.2), $\boldsymbol{\theta}^{(l)}=(b_1^{(l)},\cdots,b_p^{(l)},\boldsymbol{c}_1^{(l)},\cdots,\boldsymbol{c}_p^{(l)})^\top$ is the estimator of (l)th iteration, $\tilde{\varepsilon}_j$ is a preliminary estimation of ε_j and no need to update, $z_i=\{\mathbf{x}_i^\top,(\mathbf{x}_i\otimes(\boldsymbol{u}_i-\boldsymbol{u}_0))^\top\}^\top$. At the $(l+1)$th iteration, steps E and M are as follows:
E-step: calculate the classification probabilities $p_{ij}^{(k+1)}$,

$$p_{ij}^{(l+1)}=\frac{K_{h_0}[y_i-\sum_{k=1}^{p}\{b_k^{(l)}+\boldsymbol{c}_k^{(l)}(\boldsymbol{u}_i-\boldsymbol{u}_0)\}x_{ik}-\tilde{\varepsilon}_j]}{\sum_{j\neq i}K_{h_0}[y_i-\sum_{k=1}^{p}\{b_k^{(l)}+\boldsymbol{c}_k^{(l)}(\boldsymbol{u}_i-\boldsymbol{u}_0)\}x_{ik}-\tilde{\varepsilon}_j]}. \tag{2.8}$$

M-step: update $\boldsymbol{\theta}^{(l+1)}$

$$\begin{aligned}\boldsymbol{\theta}^{(l+1)}&=\arg\max_{\boldsymbol{\theta}}\sum_{i=1}^{n}\sum_{j\neq i}\{p_{ij}^{(l+1)}K_h(\|\boldsymbol{u}_i-\boldsymbol{u}_0\|)\log(K_{h_0}[y_i-\sum_{k=1}^{p}\{b_k+\boldsymbol{c}_k(\boldsymbol{u}_i-\boldsymbol{u}_0)\}x_{ik}-\tilde{\varepsilon}_j])\}\\&=\sum_{i=1}^{n}\sum_{j\neq i}\arg\min_{\boldsymbol{\theta}}\{p_{ij}^{(l+1)}K_h(\|\boldsymbol{u}_i-\boldsymbol{u}_0\|)[y_i-\tilde{\varepsilon}_j-z_i^\top\boldsymbol{\theta}]^2\}\\&=(\sum_{i=1}^{n}\sum_{j\neq i}p_{ij}^{(l+1)}K_h(\|\boldsymbol{u}_i-\boldsymbol{u}_0\|)z_iz_i^\top)^{-1}\sum_{i=1}^{n}\sum_{j\neq i}p_{ij}^{(l+1)}K_h(\|\boldsymbol{u}_i-\boldsymbol{u}_0\|)(y_i-\tilde{\varepsilon}_j)z_i\\&=(\boldsymbol{Z}^\top\boldsymbol{W}\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{W}\boldsymbol{Y},\end{aligned} \tag{2.9}$$

where $\boldsymbol{Z}=(\boldsymbol{Z}_{1,n-1},\cdots,\boldsymbol{Z}_{n,n-1})^\top$,

$$\boldsymbol{Z}_{i,n-1}=\begin{pmatrix}x_{i1} & x_{i1} & \cdots & x_{i1}\\ \vdots & \vdots & \vdots & \cdots\\ x_{ip} & x_{ip} & \cdots & x_{ip}\\ (\boldsymbol{u}_i-\boldsymbol{u}_0)x_{i1} & (\boldsymbol{u}_i-\boldsymbol{u}_0)x_{i1} & \cdots & (\boldsymbol{u}_i-\boldsymbol{u}_0)x_{i1}\\ \vdots & \vdots & \vdots & \cdots\\ (\boldsymbol{u}_i-\boldsymbol{u}_0)x_{ip} & (\boldsymbol{u}_i-\boldsymbol{u}_0)x_{ip} & \cdots & (\boldsymbol{u}_i-\boldsymbol{u}_0)x_{ip}\end{pmatrix}_{3p\times(n-1)},$$

$\boldsymbol{W}=diag(p_{12}^{(l+1)}K_h(\|\boldsymbol{u}_1-\boldsymbol{u}_0\|),\cdots,p_{1n}^{(l+1)}K_h(\|\boldsymbol{u}_1-\boldsymbol{u}_0\|),\cdots,p_{n,n-1}^{(l+1)}K_h(\|\boldsymbol{u}_n-\boldsymbol{u}_0\|))$, $\boldsymbol{Y}=(y_1-\tilde{\varepsilon}_2,\cdots,y_1-\tilde{\varepsilon}_n,\cdots,y_n-\tilde{\varepsilon}_{n-1})^\top$, and the second equation follows the use of Gaussian kernel. If $\|\boldsymbol{\theta}^{(l+1)}-\boldsymbol{\theta}^{(l)}\|\leq 10^{-5}$, the algorithm ends. Otherwise, the E and M steps of the algorithm continue to iterate.

Proposition 2.1. *Each iteration of the above E and M steps will monotonically increase $Q(\boldsymbol{\theta})$ in the Eq (2.6), i.e., for any l,*

$$Q(\boldsymbol{\theta}^{(l+1)}) \geq Q(\boldsymbol{\theta}^{(l)}).$$

The consistency and asymptotic of $\boldsymbol{\theta}$ are established. Let $\boldsymbol{H} = diag(1, h, h) \otimes \boldsymbol{I}_p$, where $\otimes$ is the Kronecker product and I_p is the unit matrix of $p \times p$. For $i, j = 0, 1, 2$, $k = 1, 2$, denote $\gamma_{ij} = \int u_k^i K^j(\|\boldsymbol{u}\|)d\boldsymbol{u}$ with $\boldsymbol{u} = (u_1, u_2)$, and $q(\cdot)$ is the marginal density function of $\boldsymbol{u}$.

Theorem 2.1. *Under the regularity conditions in the A, there exists a consistent maximizer $\hat{\boldsymbol{\theta}} = (\hat{b}_1, \cdots, \hat{b}_p, \hat{\boldsymbol{c}}_1, \cdots, \hat{\boldsymbol{c}}_p)^\top$ of (2.6) with probability approaching 1 such that*

$$\boldsymbol{H}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = O_p\{(nh^2)^{-1/2} + h^2\}.$$

Based on Theorem 2.1, we can know that the proposed adaptive estimator of $\boldsymbol{\theta}$ is consistent and its proof is provided in the Appendix. Next, we provide the asymptotic distribution of the proposed estimator.

Theorem 2.2. *Suppose that the regularity conditions in the A hold. Then $\hat{\boldsymbol{\theta}}$, given in Theorem 2.1, has the following asymptotic distribution*

$$\sqrt{nh^2}\{\boldsymbol{H}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \boldsymbol{S}^{-1}\frac{h^2}{2}\sum_{k=1}^{p} tr(\mathcal{H}_{\beta_k})\boldsymbol{\psi}_k(1 + o_p(1))\} \xrightarrow{D} N(\boldsymbol{0}_{3p\times 1}, [E\{\rho'(\varepsilon)^2\}]^{-1}q(\boldsymbol{u}_0)^{-1}\boldsymbol{S}^{-1}\boldsymbol{\Lambda}\boldsymbol{S}^{-1})\}\},$$

where $\rho(\cdot) = \log f(\cdot)$, $\boldsymbol{S} = diag(\gamma_{01}, \gamma_{21}, \gamma_{21}) \otimes \boldsymbol{\Gamma}(\boldsymbol{u}_0)$, $\boldsymbol{\Gamma}(\boldsymbol{u}_0) = \{\Gamma_{kj}(\boldsymbol{u}_0)\}_{1\leq k,j\leq p}$, $\Gamma_{kj}(\boldsymbol{u}_0) = E(x_{ik}x_{ij}|\boldsymbol{u}_0)$, $\boldsymbol{\Lambda} = diag(\gamma_{02}, \gamma_{22}, \gamma_{22}) \otimes \boldsymbol{\Gamma}(\boldsymbol{u}_0)$, and $\boldsymbol{\psi}_k = \binom{\gamma_{21}}{\boldsymbol{0}_{2\times 1}} \otimes (\Gamma_{kj}(\boldsymbol{u}_0))^\top_{1\leq j\leq p}$.

3. Simulation study

This section simulates the proposed adaptive estimation method and compares it with that of the local linear geographically weighted regression (LL) [24] and the Geographically Weighted Regression Model (GWR). In numerical experiments, the following four designs of error structure are considered:

1) $\varepsilon \sim N(0, 1)$;
2) $\varepsilon \sim t_3$;
3) $\varepsilon \sim 0.5N(-1, 0.5^2) + 0.5N(1, 0.5^2)$;
4) $\varepsilon \sim e^T - E(e^T)$, where $T \sim N(0, 1)$.

The first is the standard normal distribution as a benchmark for comparison, and the second is the t distribution with 3 degrees of freedom. The distributions of the third are doublet and left-biased, and the last one has a long right tail. For the above error distribution, the population positions are located at the $N = 25 \times 25$ regular grid in the square region of $D = [0, 1]^2$, and the distance between any two adjacent points in the horizontal and vertical directions is equal. At each location, the response variable $y_1, \cdots, y_n$ is generated by $y_i = \beta_1(\boldsymbol{u}_i)x_{i1} + \beta_2(\boldsymbol{u}_i)x_{i2} + \varepsilon_i$, where x_1 and x_2 follow $N(0, 1)$ with correlation coefficient $\rho = 1/\sqrt{2}$, and the regression coefficient function is as follows:

$$\beta_1(\boldsymbol{u}) = 1 + \frac{25}{12}(u_1 + u_2),$$

$$\beta_2(\boldsymbol{u}) = 1 + \frac{1}{324}[36 - (6 - \frac{25u_1}{2})^2][36 - (6 - \frac{25u_2}{2})^2],$$

the true coefficient functions contour plots of $\beta_1(\boldsymbol{u})$ and $\beta_2(\boldsymbol{u})$ are shown in Figure 1. We randomly sample $n = 200$ and 400 points from the 25×25 points in each of the 100 Monte Carlo experiments.

There are two bandwidths h and h_0 in the estimate, we use the leave-one-out cross-validation method to select h, and the choice of $h_0 = h/\log(n)$ follows Linton and Xiao [25]. The performance of the estimator $\hat{\beta}(\cdot)$ is evaluated by the square root of the average squared errors (RASE), which calculated as follows:

$$RASE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{p=1}^{2}[\hat{\beta}_p(\boldsymbol{u}_i) - \beta_p(\boldsymbol{u}_i)]^2}.$$

The simulation results are summarized in Table 1. It can be clearly seen that when the error is non-normal, the proposed adaptive estimation is better than LL and GWR, and the improvement of estimation efficiency may also be considerable. When the error is fully normally distributed, our method is still comparable to the LL and GWR method.

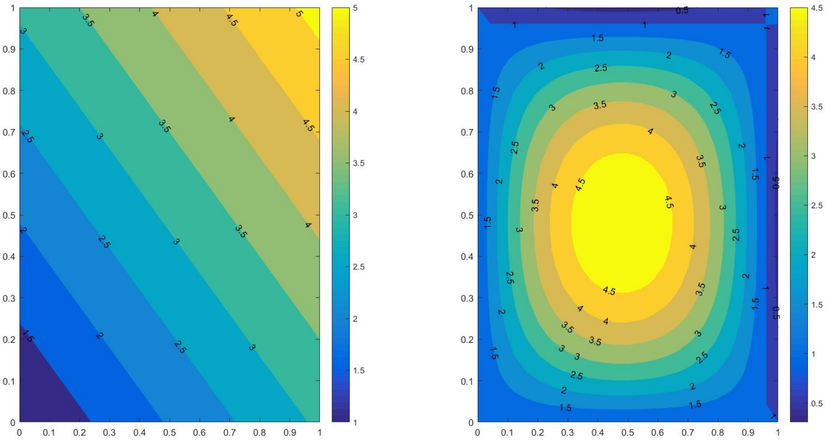


Figure 1. True coefficient functions contour plots of β_1 (left) and β_2 (right).

Table 1. Comparison RASE and its standard error in brackets.

ε	$n = 200$			$n = 400$		
	GWR	LL	Adaptive	GWR	LL	Adaptive
1	0.838(0.101)	0.787(0.099)	0.978(0.097)	0.677(0.051)	0.561(0.051)	0.790(0.048)
2	0.964(0.152)	0.857(0.123)	0.854(0.110)	0.737(0.090)	0.655(0.067)	0.538(0.028)
3	1.104(0.109)	1.009(0.107)	0.940(0.081)	0.796(0.061)	0.685(0.052)	0.632(0.031)
4	0.869(0.175)	0.837(0.127)	0.653(0.084)	0.692(0.158)	0.621(0.131)	0.405(0.060)

GWR: geographically weighted regression; LL: local linear geographically weighted regression.

Figure 2 visualizes the estimated surfaces of $\beta_1(\cdot)$ and $\beta_2(\cdot)$ using adaptive estimation method, LL and GWR based on sample size $n = 400$ when the error distribution of ε is the case 3, These results highlight that the adaptive estimation method can capture more accurate spatial pattern than the LL and GWR method.

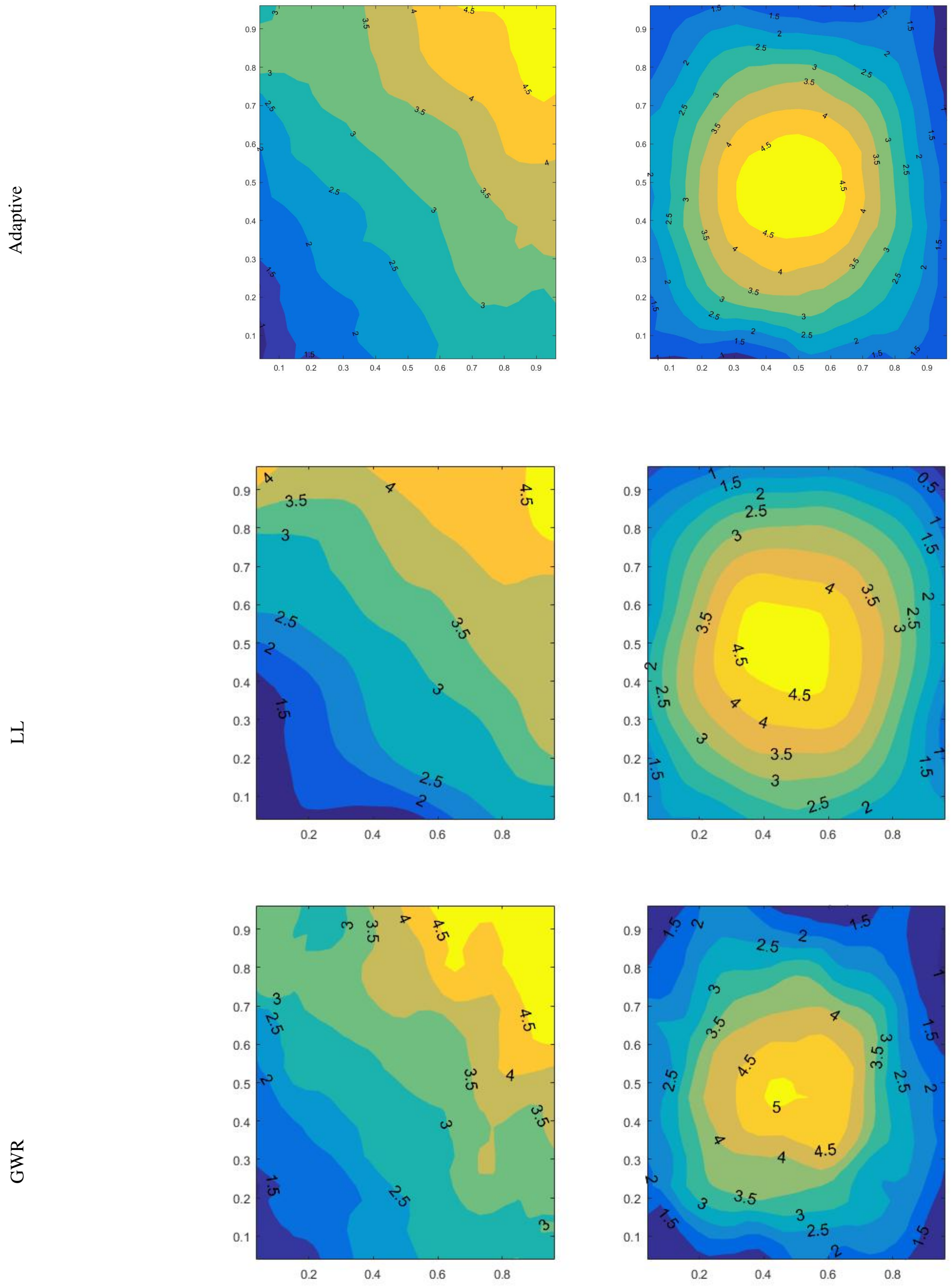
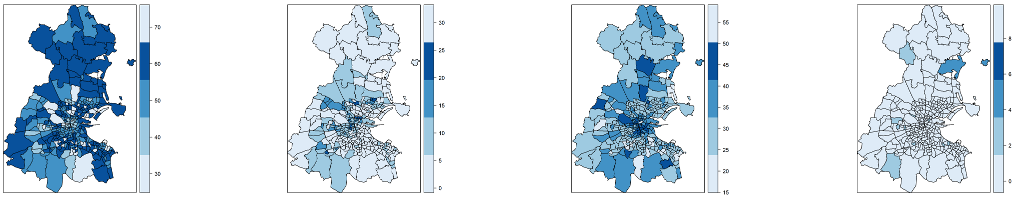


Figure 2. Estimated surface via adaptive method, LL and GWR based on sample size n = 400 when the error distribution in the case 3.

4. Real data analysis

Example 1. (Dublin Voter Turnout Data) This section applies the proposed methodology to Dublin Voter data. This dataset includes the proportion of the voting population in 322 areas, as well as several variables that may explain the change in the proportion of the voting population. Specifically, we will explore how the unemployment rate (Unempl), the proportion of ages 25 to 44 (Age 25_44) and no formal education (LowEduc) affect the proportion of the voting population in each region (GenEl2004). Figure 3 shows the spatial distribution of the dependent variable and the three independent variables.



(a) GenEl2004 (b) Unempl (c) $Age25\_44$ (d) LowEduc

Figure 3. Response and independent variables for voter turnout data in Dublin.

The dependent variable $GenEl2004$ and the independent variable $Unempl$, $Age25\_44$, $LowEduc$ are y, x_2, x_3, x_4, respectively, $x_1 = 1$ as intercept terms. We use the spatially varying coefficient models to fit the data as follows:

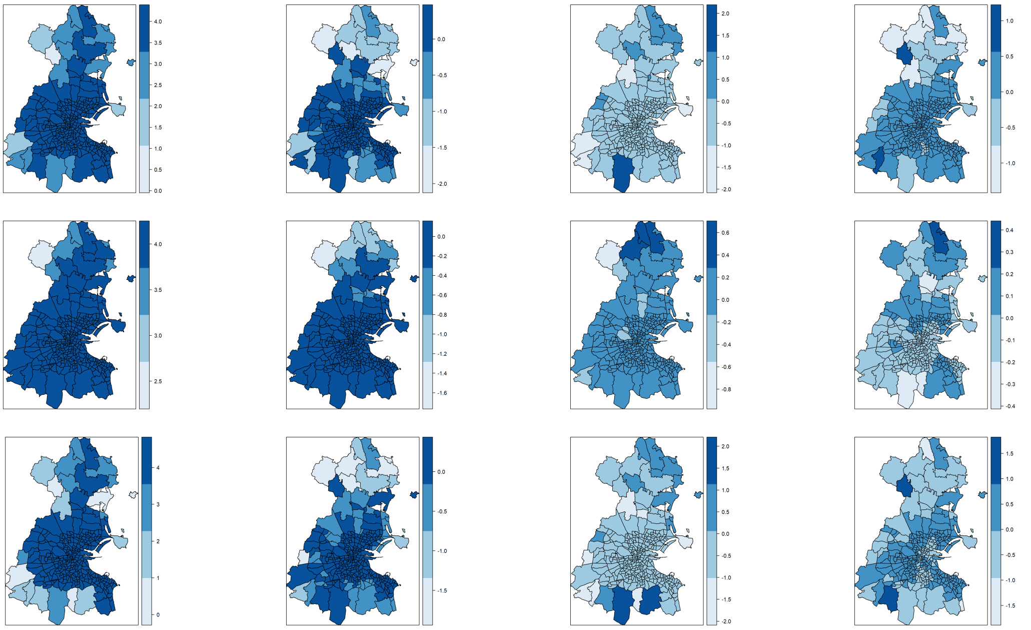
$$y_i = \beta_1(\boldsymbol{u}_i) + \sum_{k=2}^{4} \beta_p(\boldsymbol{u}_i) x_{ik} + \varepsilon_i.$$

Figure 4 summarizes estimated coefficient functions using the adaptive method, LL, and GWR respectively, which are considerably in space. Figure 7(a) shows a residual QQ-plot of the Dublin voter turnout via the adaptive method. From the plot, we can see that the distribution of the residual is very close to normal.

To evaluate the prediction accuracy of the adaptive method, we set aside 50 observations for comparing the mean squared prediction error (MSPE) of the adaptive method, LL, and GWR. The MSPE is computed as follows:

$$\text{MSPE} = \frac{1}{m} \sum_{j=1}^{m} (y_j - \hat{y}_j)^2, j = 1, \cdots, m,$$

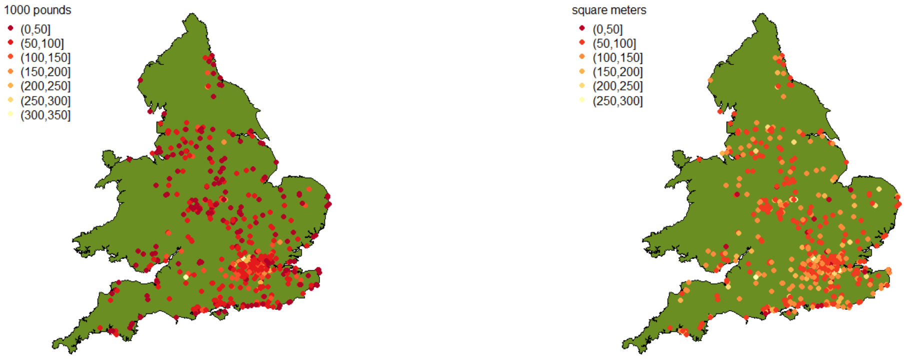
where $m = 50$ and $\hat{y}_j = \hat{\beta}_1(\boldsymbol{u}_j) + \sum_{k=2}^{4} \hat{\beta}_p(\boldsymbol{u}_j) x_{jk}$. The MSPE values by three methods are comparable, which are 0.018, 0.015 and 0.017, respectively. The QQ-plot of residuals from the Fingure 7(a) are close to the normal distribution, which explains why the MSPE of the adaptive method is very close to the MSPE of the GWR and the LL.



(a) $\hat{\beta}_1$ (b) $\hat{\beta}_2$ (c) $\hat{\beta}_3$ (d) $\hat{\beta}_4$

Figure 4. Estimated coefficient functions for voter turnout data in Dublin using adaptive method (top), LL (middle) and GWR (bottom).

Example 2. (England and Wales House Price Data) England and Wales house price data is publicly available in the R package GWmodel. The dataset includes 10 variables, namely: house sale price ($PurPrice$), $BldIntWr$, $BldPostW$, $bld60$, $bld70$, $bld80$, $TypDetch$, $TypSemiD$, $TypFlat$ and floor area ($FlrArea$). With the exception of the floor area ($FlrArea$), all independent variables are indicative variables (1 or 0). Figure 5 is shown Spatial distribution of $PurPrice$ and $FlrArea$.



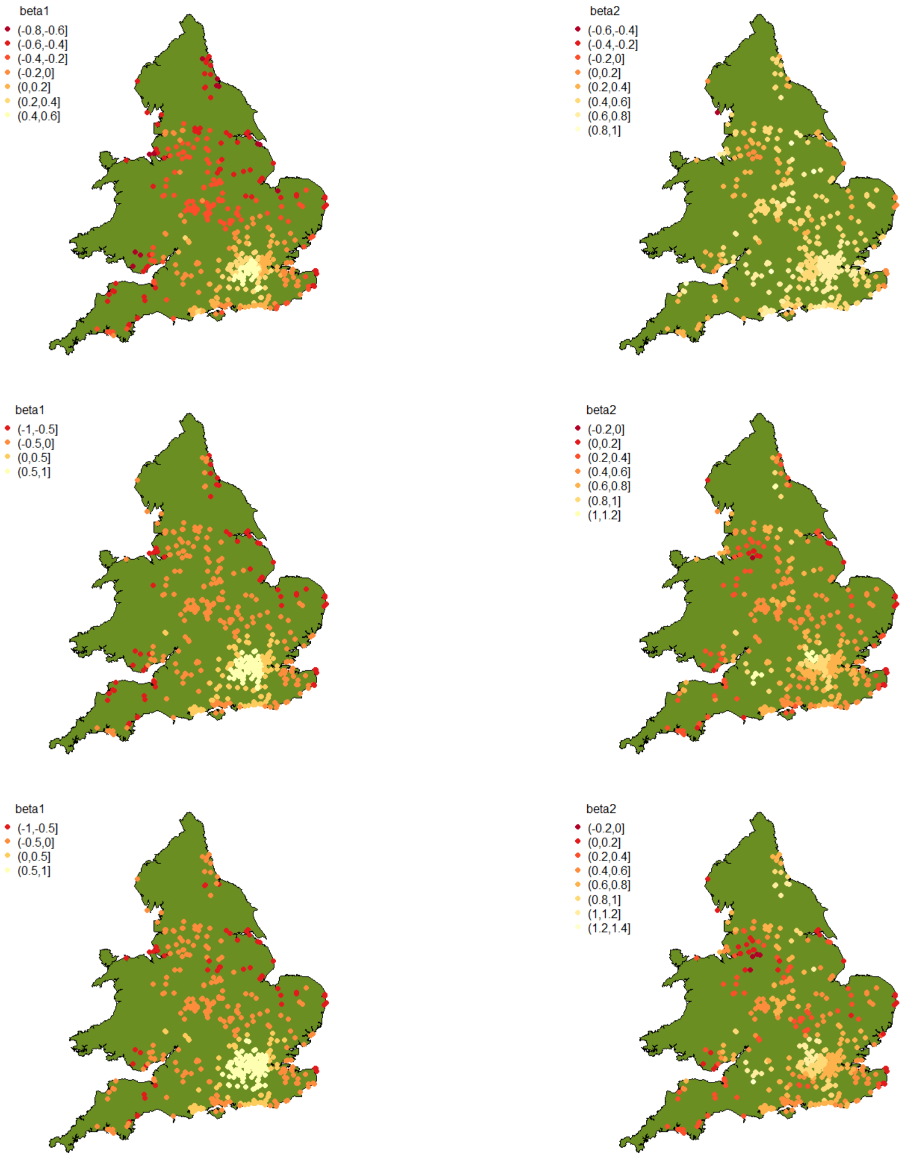
(a) Purprice (b) FlrArea

Figure 5. Response and independent variables for house price data in England and Wales.

We take the house sale price (y) as the dependent variable, FlrArea(x_2) as the independent variable, $x_1 = 1$ as the intercept term, and the spatially varying coefficient models of the fitted data is:

$$y_i = \beta_1(\boldsymbol{u}_i) + \beta_2(\boldsymbol{u}_i)x_2 + \varepsilon_i.$$

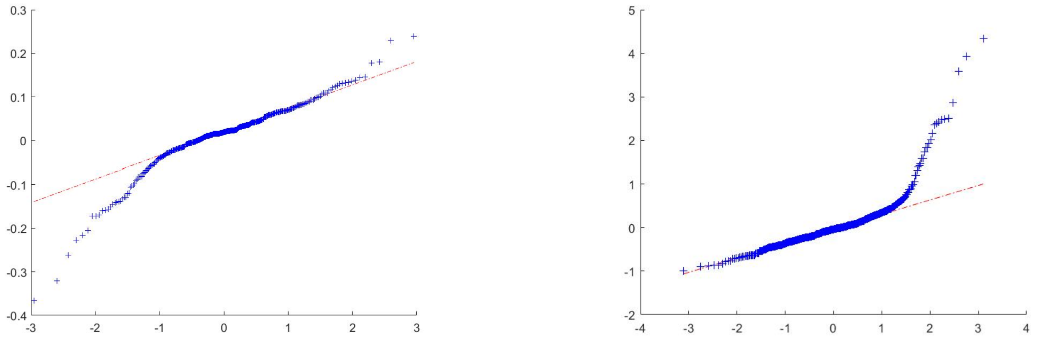
The estimated coefficient function is shown in Figure 6, and Figure 7(b) shows a residual QQ plot via adaptive method for England and Wales house price data. Similar to the analysis in example 1, we set aside 50 observations as the test set. The MSPE of the adaptive approach, LL and GWR are 0.302, 0.339 and 0.548, respectively. The QQ-plot of residuals from the above fit showed a clear deviation from normality, which explains why the MSPE from the adaptive approach is smaller than LL and GWR.



(a) $\hat{\beta}_1$

(b) $\hat{\beta}_2$

Figure 6. Estimated coefficient functions for house price data in England and Wales using adaptive method (top), LL (middle) and GWR (bottom).



(a)

(b)

Figure 7. Residual QQ-plot for two data examples: (a) Dublin voter turnout data; (b) England and Wales housing data.

5. Concluding remarks

In this article, we proposed an adaptive estimation for spatially varying coefficient models. The new estimation procedure can adapt to different errors and improve estimation efficiency than the LL and the GWR method. Simulation studies and two real data applications confirmed our theoretical findings.

The proposed method in this article can be easily extended to semiparametric varying-coefficient partially linear models, where some coefficients in the model are assumed to be constant and the remaining coefficients are allowed to spatially vary across the studied region. Another interesting future work is the spatiotemporal extension to analyze data collected across time and space.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 11871173), the National Statistical Science Research Project (Grant No. 2020LZ09).

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. C. Brunsdon, A. S. Fotheringham, M. E. Charlton, Geographically weighted regression: a method for exploring spatial nonstationarity, *Geogr. Anal.*, **28** (1996), 281–298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x
2. C. Brunsdon, A. S. Fotheringham, M. E. Charlton, Geographically weighted regression, *J. R. Stat. Soc. Ser. D-Stat.*, **47** (1998), 431–443. https://doi.org/10.1111/1467-9884.00145
3. S. L. Shen, C. L. Mei, Y. J. Zhang, Spatially varying coefficient models: testing for spatial heteroscedasticity and reweighting estimation of the coefficients, *Environ. Plann. A*, **43** (2011), 1723–1745. https://doi.org/10.1068/a43201

4. S. L. Su, C. R. Lei, A. Y. Li, J. H. Pi, Z. L. Cai, Coverage inequality and quality of volunteered geographic features in chinese cities: analyzing the associated local characteristics using geographically weighted regression, *Appl. Geogr.*, **78** (2017), 78–93. https://doi.org/10.1016/j.apgeog.2016.11.002

5. D. Al-Sulami, Z. Y. Jiang, Z. D. Lu, J. Zhu, Estimation for semiparametric nonlinear regression of irregularly located spatial time-series data, *Economet. Stat.*, **2** (2017), 22–35. https://doi.org/10.1016/j.ecosta.2017.01.002

6. Z. D. Lu, D. J. Steinskog, D. Tjøstheim, Q. W. Yao, Adaptively varying-coefficient spatiotemporal models, *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, **71** (2009), 859–880. https://doi.org/10.1111/j.1467-9868.2009.00710.x

7. Y. P. Huang, M. Yuan, Y. P. Lu, Spatially varying relationships between surface urban heat islands and driving factors across cities in China, *Environ. Plan. B-Urban*, **46** (2019), 377–394. https://doi.org/10.1177/2399808317716935

8. J. Q. Fan, W. Y. Zhang, Statistical methods with varying coefficient models, *Stat. Interface*, **1** (2008), 179–195. https://doi.org/10.4310/SII.2008.v1.n1.a15

9. T. Hastie, R. Tibshirani, Varying-coefficient models, *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, **55** (1993), 757–779. https://doi.org/10.1111/j.2517-6161.1993.tb01939.x

10. A. E. Gelfand, S. Banerjee, D. Gamerman, Spatial process modelling for univariate and multivariate dynamic spatial data, *Environmetrics*, **16** (2005), 465–479. https://doi.org/10.1002/env.715

11. R. M. Assuncao, Space varying coefficient models for small area data, *Environmetrics*, **14** (2003), 453–473. https://doi.org/10.1002/env.599

12. H. Kim, J. Lee, Hierarchical spatially varying coefficient process model, *Technometrics*, **59** (2017), 521–527. https://doi.org/10.1080/00401706.2017.1317290

13. Z. T. Luo, H. Y. Sang, B. Mallick, A Bayesian contiguous partitioning method for learning clustered latent variables, *J. Mach. Learn. Res.*, **22** (2021), 1748–1799.

14. J. R. Mu, G. N. Wang, L. Wang, Estimation and inference in spatially varying coefficient models, *Environmetrics*, **29** (2018), e2485. https://doi.org/10.1002/env.2485

15. Y. E. Shin, H. Y. Sang, D. W. Liu, T. A. Ferguson, P. X. K. Song, Autologistic network model on binary data for disease progression study, *Biometrics*, **75** (2019), 1310–1320. https://doi.org/10.1111/biom.13111

16. W. Wang, Y. Sun, Penalized local polynomial regression for spatial data, *Biometrics*, **75** (2019), 1179–1190. https://doi.org/10.1111/biom.13077

17. F. R. Li, H. Y. Sang, Spatial homogeneity pursuit of regression coefficients for large datasets, *J. Am. Stat. Assoc.*, **114** (2019), 1050–1062. https://doi.org/10.1080/01621459.2018.1529595

18. Y. Zhong, H. Y. Sang, S. J. Cook, P. M. Kellstedt, Sparse spatially clustered coefficient model via adaptive regularization, *Comput. Stat. Data Anal.*, **177** (2023), 107581. https://doi.org/10.1016/j.csda.2022.107581

19. C. Stein, *Efficient nonparametric testing and estimation*, University California Press, 1956.

20. Y. X. Chen, Q. Wang, W. X. Yao, Adaptive estimation for varying coefficient models, *J. Multivar. Anal.*, **137** (2015), 17–31. https://doi.org/10.1016/j.jmva.2015.01.017

21. Z. Y. Zhou, J. Yu, Adaptive estimation for varying coefficient models with non stationary covariates, *Commun. Stat. Theory M.*, **48** (2019), 4034–4050. https://doi.org/10.1080/03610926.2018.1484483
22. W. X. Yao, A note on EM algorithm for mixture models, *Stat. Probabil. Lett.*, **83** (2013), 519–526. https://doi.org/10.1016/j.spl.2012.10.017
23. L. Jia, S. Ray, B. G. Lindsay, A nonparametric statistical approach to clustering via mode identification, *J. Mach. Learn. Res.*, **8** (2007), 1687–1723.
24. N. Wang, C. L. Mei, X. D. Yan, Local linear estimation of spatially varying coefficient models: an improvement on the geographically weighted regression technique, *Environ. Plann. A*, **40** (2008), 986–1005. https://doi.org/10.1068/a3941
25. O. Linton, Z. J. Xiao, A nonparametric regression estimator that adapts to error distribution of unknown form, *Economet. Theory*, **23** (2007), 371–413. https://doi.org/10.1017/S026646660707017X

A. Appendix

This section will give proofs of propositions 2.1, theorem 2.1 and theorem 2.2, with the required regular conditions as follows:

1) $K(\cdot)$ is bounded, symmetric, and has bounded support and bounded derivatives;
2) $\{\mathbf{x}_i\}_{i=1}^n$, $\{\boldsymbol{u}_i\}_{i=1}^n$, $\{\varepsilon_i\}_{i=1}^n$ are independent and identically distributed and $\{\varepsilon_i\}_{i=1}^n$ is independent of $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\boldsymbol{u}_i\}_{i=1}^n$. In addition, the independent variable $\mathbf{x}$ has bounded support;
3) The probability density function $f(\varepsilon)$ of ε has fourth-order bounded continuous derivative. Assume $E[\rho'(\varepsilon)] = 0$, $E[\rho''(\varepsilon)] < \infty$, $E[\rho'(\varepsilon)^2] < \infty$ and $\rho'''(\cdot)$ is bounded;
4) The marginal density $q(\boldsymbol{u})$ of $\boldsymbol{u}$ has a continuous second derivative in some neighborhood of $\boldsymbol{u}_0$ and $q(\boldsymbol{u}_0) \neq 0$;
5) $h \to 0$, when $n \to \infty$, $nh \to \infty$, $h_0 = h/\log(n)$;
6) $\beta_k(\cdot), k = 1, \cdots, p$ has bounded and continuous third derivative.

Proof of propositions 2.1: Note that

$$
\begin{aligned}
& Q(\boldsymbol{\theta}^{(l+1)}) - Q(\boldsymbol{\theta}^{(l)}) \\
&= \sum_{i=1}^{n} K_h(\|\boldsymbol{u}_i - \boldsymbol{u}_0\|) \log \left\{ \frac{\sum_{j\neq i} K_{h_0}\left[y_i - \sum_{k=1}^{p}\left\{b_k^{(l+1)} + \boldsymbol{c}_k^{(l+1)}(\boldsymbol{u}_i - \boldsymbol{u}_0)\right\} x_{ik} - \tilde{\varepsilon}_j\right]}{\sum_{j\neq i} K_{h_0}\left[y_i - \sum_{k=1}^{p}\left\{b_k^{(l)} + \boldsymbol{c}_k^{(l)}(\boldsymbol{u}_i - \boldsymbol{u}_0)\right\} x_{ik} - \tilde{\varepsilon}_j\right]} \right\} \\
&= \sum_{i=1}^{n} K_h(\|\boldsymbol{u}_i - \boldsymbol{u}_0\|) \log \sum_{j\neq i} \left(\frac{K_{h_0}\left[y_i - \sum_{k=1}^{p}\left\{b_k^{(l)} + \boldsymbol{c}_k^{(l)}(\boldsymbol{u}_i - \boldsymbol{u}_0)\right\} x_{ik} - \tilde{\varepsilon}_j\right]}{\sum_{j\neq i} K_{h_0}\left[y_i - \sum_{k=1}^{p}\left\{b_k^{(l)} + \boldsymbol{c}_k^{(l)}(\boldsymbol{u}_i - \boldsymbol{u}_0)\right\} x_{ik} - \tilde{\varepsilon}_j\right]} \right) \\
&\quad \times \left(\frac{K_{h_0}\left[y_i - \sum_{k=1}^{p}\left\{b_k^{(l+1)} + \boldsymbol{c}_k^{(l+1)}(\boldsymbol{u}_i - \boldsymbol{u}_0)\right\} x_{ik} - \tilde{\varepsilon}_j\right]}{K_{h_0}\left[y_i - \sum_{k=1}^{p}\left\{b_k^{(l)} + \boldsymbol{c}_k^{(l)}(\boldsymbol{u}_i - \boldsymbol{u}_0)\right\} x_{ik} - \tilde{\varepsilon}_j\right]} \right) \\
&= \sum_{i=1}^{n} K_h(\|\boldsymbol{u}_i - \boldsymbol{u}_0\|) \log \left\{ \sum_{j\neq i} p_{ij}^{(l+1)} \frac{K_{h_0}\left[y_i - \sum_{k=1}^{p}\left\{b_k^{(l+1)} + \boldsymbol{c}_k^{(l+1)}(\boldsymbol{u}_i - \boldsymbol{u}_0)\right\} x_{ik} - \tilde{\varepsilon}_j\right]}{K_{h_0}\left[y_i - \sum_{k=1}^{p}\left\{b_k^{(l)} + \boldsymbol{c}_k^{(l)}(\boldsymbol{u}_i - \boldsymbol{u}_0)\right\} x_{ik} - \tilde{\varepsilon}_j\right]} \right\},
\end{aligned}
$$

where

$$
p_{ij}^{(l+1)} = \frac{K_{h_0}\left[y_i - \sum_{k=1}^{p}\left\{b_k^{(l)} + \boldsymbol{c}_k^{(l)}(\boldsymbol{u}_i - \boldsymbol{u}_0)\right\} x_{ik} - \tilde{\varepsilon}_j\right]}{\sum_{j\neq i} K_{h_0}\left[y_i - \sum_{k=1}^{p}\left\{b_k^{(l)} + \boldsymbol{c}_k^{(l)}(\boldsymbol{u}_i - \boldsymbol{u}_0)\right\} x_{ik} - \tilde{\varepsilon}_j\right]}.
$$

From Jensen's inequality,

$$
\begin{aligned}
& Q(\boldsymbol{\theta}^{(k+1)}) - Q(\boldsymbol{\theta}^{(k)}) \\
&\geq \sum_{i=1}^{n} K_h(\|\boldsymbol{u}_i - \boldsymbol{u}_0\|) \sum_{j\neq i} p_{ij}^{(l+1)} \log \left\{ \frac{K_{h_0}\left[y_i - \sum_{k=1}^{p}\left\{b_k^{(l+1)} + \boldsymbol{c}_k^{(l+1)}(\boldsymbol{u}_i - \boldsymbol{u}_0)\right\} x_{ik} - \tilde{\varepsilon}_j\right]}{K_{h_0}\left[y_i - \sum_{k=1}^{p}\left\{b_k^{(l)} + \boldsymbol{c}_k^{(l)}(\boldsymbol{u}_i - \boldsymbol{u}_0)\right\} x_{ik} - \tilde{\varepsilon}_j\right]} \right\}.
\end{aligned}
$$

Based on the properties of step M of formula (2.9), it is proved that $Q(\boldsymbol{\theta}^{(k+1)}) - Q(\boldsymbol{\theta}^{(k)}) \geq 0$.

Proof of theorem 2.1: According to the the result of Linton and Xiao [25], the asymptotic behaviour of $\hat{\boldsymbol{\theta}}$ in (7) is the same as that obtained from (4). Therefore, we prove the asymptotic properties of $\hat{\boldsymbol{\theta}}$ based on (2.3).

Denote $\boldsymbol{\theta}^* = \boldsymbol{H}\boldsymbol{\theta}$, $\boldsymbol{x}_i^* = \left(x_{i1}, \cdots, x_{ip}, (\frac{\boldsymbol{u}_i - \boldsymbol{u}_0}{h})^\top x_{i1}, \cdots, (\frac{\boldsymbol{u}_i - \boldsymbol{u}_0}{h})^\top x_{ip}\right)^\top$, $K_i = K_h(\|\boldsymbol{u}_i - \boldsymbol{u}_0\|)$, $R(\boldsymbol{u}_i, \boldsymbol{x}_i) = \sum_{k=1}^{p} \beta_k(\boldsymbol{u}_i) x_{ik} - \sum_{k=1}^{p}\{b_k + \boldsymbol{c}_k(\boldsymbol{u}_i - \boldsymbol{u}_0)\}x_{ik}$, and $a_n = (nh^2)^{-1/2} + h^2$. Let $\rho(\cdot) = \log f(\cdot)$, objective function (2.3) is written as

$$
L(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} K_i \rho(y_i - \boldsymbol{\theta}^{*\top} \boldsymbol{x}_i^*) \triangleq L(\boldsymbol{\theta}^*).
$$

Based on the definition of $\boldsymbol{\theta}^*$, it is sufficient to show that for any given $\eta > 0$, there exists a large constant c such that

$$
P\{\sup_{\|\mu\|=1} L(\boldsymbol{\theta}^* + a_n \mu) < L(\boldsymbol{\theta}^*)\} \geq 1 - \eta
$$

where μ has the same dimension as $\boldsymbol{\theta}$, a_n is the convergence rate. By using Taylor expansion, it follows that

$$\begin{aligned}L(\boldsymbol{\theta}^* + a_n\mu) - L(\boldsymbol{\theta}^*) &= \frac{1}{n}\sum_{i=1}^{n} K_i\{\rho(\varepsilon_i + R(\boldsymbol{u}_i, \boldsymbol{x}_i) - a_n\mu^\top \boldsymbol{x}_i^*) - \rho(\varepsilon_i + R(\boldsymbol{u}_i, \boldsymbol{x}_i))\}\\ &= -\frac{1}{n}\sum_{i=1}^{n} K_i\rho'(\varepsilon_i + R(\boldsymbol{u}_i, \boldsymbol{x}_i))a_n\mu^\top \boldsymbol{x}_i^* + \frac{1}{2n}\sum_{i=1}^{n} K_i\rho''(\varepsilon_i + R(\boldsymbol{u}_i, \boldsymbol{x}_i))a_n^2(\mu^\top \boldsymbol{x}_i^*)^2\\ &\quad - \frac{1}{6n}\sum_{i=1}^{n} K_i\rho'''(z_i)a_n^3(\mu^\top \boldsymbol{x}_i^*)^3\\ &\triangleq I_1 + I_2 + I_3,\end{aligned}$$

where z_i is a value between $\varepsilon_i + R(\boldsymbol{u}_i, \boldsymbol{x}_i) - a_n\mu^\top \boldsymbol{x}_i^*$ and $\varepsilon_i + R(\boldsymbol{u}_i, \boldsymbol{x}_i)$.

For $I_1 = -\frac{1}{n}\sum_{i=1}^{n} K_i\rho'(\varepsilon_i + R(\boldsymbol{u}_i, \boldsymbol{x}_i))a_n\mu^\top \boldsymbol{x}_i^*$, Let $\delta_1 = E[\rho''(\varepsilon_i)]$. Since $R(\boldsymbol{u}_i, \boldsymbol{x}_i) = \sum_{k=0}^{p}\beta_k(\boldsymbol{u}_i)x_{ik} - \sum_{k=0}^{p}\{b_k + \boldsymbol{c}_k(\boldsymbol{u}_i - \boldsymbol{u}_0)\}x_{ik} = O_p(h^2) = o_p(1)$ and $E[\rho'(\varepsilon)] = 0$, so

$$\begin{aligned}E(I_1) &= -E(K_i\rho'(\varepsilon_i + R(\boldsymbol{u}_i, \boldsymbol{x}_i))a_n\mu^\top \boldsymbol{x}_i^*)\\ &\approx -a_nE\{K_i\rho''(\varepsilon_i)R(\boldsymbol{u}_i, \boldsymbol{x}_i)\mu^\top \boldsymbol{x}_i^*\}\\ &= -a_nE[\rho''(\varepsilon_i)]E[K_iR(\boldsymbol{u}_i, \boldsymbol{x}_i)\mu^\top \boldsymbol{x}_i^*]\\ &= -a_n\delta_1E[K_iR(\boldsymbol{u}_i, \boldsymbol{x}_i)\mu^\top \boldsymbol{x}_i^*]\\ &= -a_n\delta_1E\{E[R(\boldsymbol{u}_i, \boldsymbol{x}_i)\mu^\top \boldsymbol{x}_i^*|\boldsymbol{u}_i]K_i\}\end{aligned}$$

By using $\mu^\top \boldsymbol{x}_i^* \le ||\mu|| \cdot ||\boldsymbol{x}_i^*||$, we have $E(I_1) = O(a_nh^2)$.

$$\begin{aligned}var(I_1) &= \frac{1}{n}var\{K_i\rho'(\varepsilon_i + R(\boldsymbol{u}_i, \boldsymbol{x}_i))a_n\mu^\top \boldsymbol{x}_i^*\}\\ &= \frac{1}{n}\{E(A^2) - [E(A)]^2\}\end{aligned}$$

where $A = K_i\rho'(\varepsilon_i + R(\boldsymbol{u}_i, \boldsymbol{x}_i))a_n\mu^\top \boldsymbol{x}_i^*$. Let $\delta_2 = E[\rho'(\varepsilon_i)^2]$, then

$$\begin{aligned}E(A^2) &= E\{K_i^2\rho'(\varepsilon_i + R(\boldsymbol{u}_i, \boldsymbol{x}_i))^2a_n^2(\mu^\top \boldsymbol{x}_i^*)^2\}\\ &\approx a_n^2E\{K_i^2\rho'(\varepsilon_i)^2(\mu^\top \boldsymbol{x}_i^*)^2\}\\ &= a_n^2\delta_2E\{E[\mu^\top \boldsymbol{x}_i^*\boldsymbol{x}_i^{*\top}\mu|\boldsymbol{u}_i]K_i^2\}\\ &= a_n^2\delta_2\mu^\top E\{E[\boldsymbol{x}_i^*\boldsymbol{x}_i^{*\top}|\boldsymbol{u}_i]K_i^2\}\mu\end{aligned}$$

Note that $\boldsymbol{x}_i^*\boldsymbol{x}_i^{*\top} = \left(x_{ij}x_{ik}(\frac{\boldsymbol{u}_i-\boldsymbol{u}_0}{h})^l((\frac{\boldsymbol{u}_i-\boldsymbol{u}_0}{h})^{l'})^\top\right)_{1\le j,k\le p;l,l'=0,1}$, $\Gamma_{jk}(\boldsymbol{u}_i) = E(x_{ij}x_{ik}|\boldsymbol{u}_i)$, $\int_{R^2}\boldsymbol{u}K(||\boldsymbol{u}||)d\boldsymbol{u} = \boldsymbol{0}_{2\times1}$,

$\int_{R^2} \boldsymbol{u}\boldsymbol{u}'\boldsymbol{u}K(\|\boldsymbol{u}\|)d\boldsymbol{u} = \mathbf{0}_{2\times1}$, $\int_{R^2} u_1u_2K(\|\boldsymbol{u}\|)d\boldsymbol{u} = \mathbf{0}_{2\times1}$, for $1 \le j, k \le p$, then

$$\begin{aligned}
&E\left[E(x_{ij}x_{ik}|\boldsymbol{u}_i)\left(\frac{\boldsymbol{u}_i-\boldsymbol{u}_0}{h}\right)^l\left(\left(\frac{\boldsymbol{u}_i-\boldsymbol{u}_0}{h}\right)^{l'}\right)^\top K_i^2\right]\\
&=E\left[\Gamma_{jk}(\boldsymbol{u}_i)\left(\frac{\boldsymbol{u}_i-\boldsymbol{u}_0}{h}\right)^l\left(\left(\frac{\boldsymbol{u}_i-\boldsymbol{u}_0}{h}\right)^{l'}\right)^\top K_i^2\right]\\
&=\frac{1}{h^4}\int\Gamma_{jk}(\boldsymbol{u}_i)\left(\frac{\boldsymbol{u}_i-\boldsymbol{u}_0}{h}\right)^l\left(\left(\frac{\boldsymbol{u}_i-\boldsymbol{u}_0}{h}\right)^{l'}\right)^\top K^2\left(\frac{\|\boldsymbol{u}_i-\boldsymbol{u}_0\|}{h}\right)q(\boldsymbol{u}_i)d\boldsymbol{u}_i\\
&=\frac{1}{h^2}q(\boldsymbol{u}_0)\Gamma_{jk}(\boldsymbol{u}_0)\int\boldsymbol{t}^l(\boldsymbol{t}^{l'})^\top K^2(\|\boldsymbol{t}\|)d\boldsymbol{t}
\end{aligned}\tag{A.1}$$

The second equation follows the Taylor expansion, and the assumption ε is independent of $\boldsymbol{u}$ and $\boldsymbol{x}$. Then, $E\{E[\boldsymbol{x}_i^*\boldsymbol{x}_i^{*\top}|\boldsymbol{u}_i]K_i^2\} = \frac{1}{h^2}q(\boldsymbol{u}_0)\boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda} = diag(1, \nu_2, \nu_2)\otimes\boldsymbol{\Gamma}(\boldsymbol{u}_0)$ is a $3p\times3p$ matrix. Thus,

$$E(A^2) = a_n^2\delta_2\frac{1}{h}q(\boldsymbol{u}_0)\mu^\top\boldsymbol{\Lambda}\mu = O(a_n^2\frac{1}{h^2}).$$

Note that $[E(A)]^2 = [E(I_1)]^2 = [O(a_nh^2]^2 \ll E(A^2)$, then $var(I_1) \approx \frac{1}{n}[E(A)]^2 = O(a_n^2\frac{1}{nh^2})$. Hence,

$$I_1 = E(I_1) + O_p(\sqrt{var(I_1)}) = O_p(a_nh^2) + O_p(\sqrt{a_n^2\frac{1}{nh^2}}) = O_p(a_n^2).$$

Similarly,

$$I_2 = \frac{1}{2n}\sum_{i=1}^n K_i\rho''(\varepsilon_i + R(\boldsymbol{u}_i, \boldsymbol{x}_i))a_n^2(\mu^\top\boldsymbol{x}_i^*)^2 = O_p(a_n^2),$$

and

$$I_3 = \frac{1}{6n}\sum_{i=1}^n K_i\rho'''(z_i)a_n^3(\mu^\top\boldsymbol{x}_i^*)^3 = O_p(a_n^3).$$

Assume $\delta_1 > 0$, we can choose c large enough such that $I_1 + I_2 + I_3 < 0$ with probability at least $1-\eta$. Thus $P\left\{\sup_{\|\mu\|=c} L(\boldsymbol{\theta}^* + a_n\mu) < L(\boldsymbol{\theta}^*)\right\} \ge 1-\eta$.

Proof of theorem 2.2: Since $\hat{\boldsymbol{\theta}}^*$ maximizes $L(\boldsymbol{\theta}^*)$, then $L'(\hat{\boldsymbol{\theta}}^*) = 0$. By Taylor expansion,

$$0 = L'(\hat{\boldsymbol{\theta}}^*) = L'(\boldsymbol{\theta}^*) + L''(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*) + o_p(1),$$

thus

$$\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^* = -[L''(\boldsymbol{\theta}^*)]^{-1}L'(\boldsymbol{\theta}^*)(1 + o_p(1)).$$

For $L''(\boldsymbol{\theta}^*)$, since $L(\boldsymbol{\theta}^*) = \frac{1}{n}\sum_{i=1}^n K_i\rho(y_i - \boldsymbol{\theta}^{*\top}\boldsymbol{x}_i^*)$, and $y_i - \boldsymbol{\theta}^{*\top}\boldsymbol{x}_i^* = \varepsilon_i + R(\boldsymbol{u}_i, \boldsymbol{x}_i)$, then $L''(\boldsymbol{\theta}^*) = \frac{1}{n}\sum_{i=1}^n K_i\rho''(\varepsilon_i + R(\boldsymbol{u}_i, \boldsymbol{x}_i))\boldsymbol{x}_i^*\boldsymbol{x}_i^{*\top}$, and the expectation is

$$\begin{aligned}
E[L''(\boldsymbol{\theta}^*)] &= E\{K_i\rho''(\varepsilon_i + R(\boldsymbol{u}_i, \boldsymbol{x}_i))\boldsymbol{x}_i^*\boldsymbol{x}_i^{*\top}\}\\
&\approx E\{K_i\rho''(\varepsilon_i)\boldsymbol{x}_i^*\boldsymbol{x}_i^{*\top}\}\\
&= \delta_1E\{E\{\boldsymbol{x}_i^*\boldsymbol{x}_i^{*\top}|\boldsymbol{u}_i\}K_i\}\\
&= \delta_1q(\boldsymbol{u}_0)\boldsymbol{Q}(1 + o(1)),
\end{aligned}$$

where $\boldsymbol{S} = diag(\gamma_{01}, \gamma_{21}, \gamma_{21}) \otimes \boldsymbol{\Gamma}(\boldsymbol{u}_0)$, the last equation follow (A.1). In this study, we consider the element-wise variance of a matrix, then

$$\begin{aligned} var[L''(\boldsymbol{\theta}^*)] &= \frac{1}{n} var\{K_i \rho''(\varepsilon_i + R(\boldsymbol{u}_i, \boldsymbol{x}_i))\boldsymbol{x}_i^* \boldsymbol{x}_i^{*\top}\} \\ &= O_p(\frac{1}{nh^2}). \end{aligned}$$

Based on the result $L''(\boldsymbol{\theta}^*) = E[L''(\boldsymbol{\theta}^*)] + O_p\sqrt{var[L''(\boldsymbol{\theta}^*)]}$ and the assumption $nh \to \infty$, it follows that $L''(\boldsymbol{\theta}^*) = \delta_1 q(\boldsymbol{u}_0)\boldsymbol{S}(1 + o_p(1))$.

For $L'(\boldsymbol{\theta}^*)$,

$$\begin{aligned} L'(\boldsymbol{\theta}^*) &= -\frac{1}{n}\sum_{i=1}^{n} K_i \rho'(\varepsilon_i + R(\boldsymbol{u}_i, \boldsymbol{x}_i))\boldsymbol{x}_i^* \\ &= -\frac{1}{n}\sum_{i=1}^{n} K_i \rho'(\varepsilon_i)\boldsymbol{x}_i^* - \frac{1}{n}\sum_{i=1}^{n} K_i \rho''(\varepsilon_i) R(\boldsymbol{u}_i, \boldsymbol{x}_i))\boldsymbol{x}_i^* \\ &\triangleq -\boldsymbol{w}_m - \boldsymbol{v}_n \end{aligned}$$

The asymptotic result is determined by $\boldsymbol{w}_m$. Next, calculating the order of $\boldsymbol{v}_n$.

$$E(\boldsymbol{v}_n) = E[K_i \rho''(\varepsilon_i) R(\boldsymbol{u}_i, \boldsymbol{x}_i))x_i^*] = \delta_1 E\{E\{R(\boldsymbol{u}_i, \boldsymbol{x}_i)x_i^* | \boldsymbol{u}_i\} K_i\}$$

For $R(\boldsymbol{u}_i, \boldsymbol{x}_i)x_i^*$, since $\beta_j'''(\cdot)$ is bounded, then we have

$$\begin{aligned} R(\boldsymbol{u}_i, \boldsymbol{x}_i) &= \sum_{k=1}^{p} \beta_k(\boldsymbol{u}_i) x_{ik} - \sum_{k=1}^{p} \{b_k + \boldsymbol{c}_k(\boldsymbol{u}_i - \boldsymbol{u}_0)\} x_{ik} \\ &= \sum_{k=1}^{p} \frac{1}{2}(\boldsymbol{u}_i - \boldsymbol{u}_0)^\top \mathcal{H}_{\beta_k}(\boldsymbol{u}_0)(\boldsymbol{u}_i - \boldsymbol{u}_0) x_{ik}(1 + o_p(1)) \end{aligned}$$

where $\mathcal{H}_{\beta_k}$ is Hessian matrix. By $\boldsymbol{x}_i^* = (x_{i1}, \cdots, x_{ip}, (\frac{\boldsymbol{u}_i - \boldsymbol{u}_0}{h})^\top x_{i1}, \cdots, (\frac{\boldsymbol{u}_i - \boldsymbol{u}_0}{h})^\top x_{ip})^\top$,

$$\begin{aligned} R(\boldsymbol{u}_i, \boldsymbol{x}_i)\boldsymbol{x}_i^* = \Bigg\{ & \left(\sum_{k=1}^{p} \frac{1}{2}(\boldsymbol{u}_i - \boldsymbol{u}_0)^\top \mathcal{H}_{\beta_k}(\boldsymbol{u}_0)(\boldsymbol{u}_i - \boldsymbol{u}_0) x_{ik} x_{ij} \right)_{1 \le j \le p}, \\ & \left(\sum_{k=1}^{p} \frac{1}{2h}(\boldsymbol{u}_i - \boldsymbol{u}_0)^\top \mathcal{H}_{\beta_k}(\boldsymbol{u}_0)(\boldsymbol{u}_i - \boldsymbol{u}_0)(\boldsymbol{u}_i - \boldsymbol{u}_0)^\top x_{ik} x_{ij} \right)_{1 \le j \le p} \Bigg\}^\top_{3p \times 1}. \end{aligned}$$

The expectation about the first item is

$$
\begin{aligned}
&E\left\{E\left[\sum_{k=1}^{p}\frac{1}{2}(\boldsymbol{u}_i-\boldsymbol{u}_0)^{\top}\mathcal{H}_{\beta_k}(\boldsymbol{u}_0)(\boldsymbol{u}_i-\boldsymbol{u}_0)x_{ik}x_{ij}|\boldsymbol{u}_i\right]K_i\right\}\\
&=E\left\{\sum_{k=1}^{p}\frac{1}{2}(\boldsymbol{u}_i-\boldsymbol{u}_0)^{\top}\mathcal{H}_{\beta_k}(\boldsymbol{u}_0)(\boldsymbol{u}_i-\boldsymbol{u}_0)\Gamma_{kj}(\boldsymbol{u}_i)K_i\right\}\\
&=\frac{h^2}{2}q(\boldsymbol{u}_0)\sum_{k=1}^{p}\Gamma_{kj}(\boldsymbol{u}_0)\int\boldsymbol{t}^{\top}\mathcal{H}_{\beta_k}\boldsymbol{t}K(\|\boldsymbol{t}\|)d\boldsymbol{t}\\
&=\frac{h^2}{2}q(\boldsymbol{u}_0)\sum_{k=1}^{p}tr(\mathcal{H}_{\beta_k})\Gamma_{kj}(\boldsymbol{u}_0)\gamma_{21},
\end{aligned}
$$

and the expectation on second item is

$$
\begin{aligned}
&E\left\{E\left[\sum_{k=1}^{p}\frac{1}{2h}(\boldsymbol{u}_i-\boldsymbol{u}_0)^{\top}\mathcal{H}_{\beta_k}(\boldsymbol{u}_0)(\boldsymbol{u}_i-\boldsymbol{u}_0)^2x_{ik}x_{ij}|\boldsymbol{u}_i\right]K_i\right\}\\
&=\frac{h^2}{2}q(\boldsymbol{u}_0)\sum_{k=1}^{p}\Gamma_{kj}(\boldsymbol{u}_0)\int\boldsymbol{t}^{\top}\mathcal{H}_{\beta_k}\boldsymbol{t}\boldsymbol{t}^{\top}K(\|\boldsymbol{t}\|)\mathrm{d}\boldsymbol{t}\\
&=\boldsymbol{0}_{2\times1},
\end{aligned}
$$

then

$$E(\boldsymbol{\nu}_n)=\delta_1\frac{h^2}{2}q(\boldsymbol{u}_0)\sum_{k=1}^{p}tr(\mathcal{H}_{\beta_k})\boldsymbol{\psi}_j(1+o(1))$$

where $\boldsymbol{\psi}_k=\binom{\gamma_{21}}{\boldsymbol{0}_{2\times1}}\otimes(\Gamma_{kj}(\boldsymbol{u}_0))^{\top}_{1\le j\le p}$ is a $3p\times1$ vector for $j=1,\cdots,p$. Since $var(\boldsymbol{\nu}_n)=\frac{1}{n}var\{K_i\rho''(\varepsilon_i)R(\boldsymbol{u}_i,\boldsymbol{x}_i))\boldsymbol{x}_i^*\}=O(h^2/n)$, then based on the result $\boldsymbol{\nu}_n=E(\boldsymbol{\nu}_n)+var(\sqrt{\boldsymbol{\nu}_n})$ and the assumption $nh\to\infty$, it follows that

$$\boldsymbol{\nu}_n=\delta_1\frac{h^2}{2}q(\boldsymbol{u}_0)\sum_{k=1}^{p}tr(\mathcal{H}_{\beta_k})\boldsymbol{\psi}_j(1+o_p(1)).$$

Then

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}^*-\boldsymbol{\theta}^*&=-[L''(\boldsymbol{\theta}^*)]^{-1}L'(\boldsymbol{\theta}^*)(1+o_p(1))\\
&=\frac{\boldsymbol{S}^{-1}\boldsymbol{w}_n}{\delta q(\boldsymbol{u}_0)}(1+o_p(1))+\boldsymbol{S}^{-1}\frac{h^2}{2}\sum_{k=1}^{p}tr(\mathcal{H}_{\beta_k})\boldsymbol{\psi}_j(1+o_p(1)).
\end{aligned}
$$

For $\boldsymbol{w}_n$, based on the assumption $E[\rho'(\varepsilon_i)]=0$, we can easily get $E(\boldsymbol{w}_n)=0$, and

$$
\begin{aligned}
var(\boldsymbol{w}_n)&=\frac{1}{n}var\{\frac{1}{n}K_i\rho'(\varepsilon_i)x_i^*\}\\
&=\frac{1}{n}E\{K_i^2\rho'(\varepsilon_i)^2\boldsymbol{x}_i^*\boldsymbol{x}_i^{*\top}\}\\
&=\frac{1}{nh^2}\delta_2q(\boldsymbol{u}_0)\boldsymbol{\Lambda}(1+o(1)).
\end{aligned}
$$

Based on Lyapunov Central Limit Theorem, we have the following result

$$\sqrt{nh^2}\{\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^* - \boldsymbol{S}^{-1}\frac{h^2}{2}\sum_{k=1}^{p} tr(\mathcal{H}_{\beta_k})\boldsymbol{\psi}_k(1 + o_p(1))\} \xrightarrow{D} N(\mathbf{0}_{3p\times 1}, \delta_1^{-2}\delta_2 q(\boldsymbol{u}_0)^{-1}S^{-1}\mathbf{\Lambda}S^{-1}).$$

By $\delta_1^{-1} = \delta_2$, the theorem is proved.