Wang Wu ORCID iD: 0000-0003-2633-1441

Penalized Local Polynomial Regression for Spatial Data

Wu Wang* and Ying Sun**

King Abdullah University of Science and Technology (KAUST),
Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE),
Thuwal 23955-6900, Saudi Arabia

**email:* wu.wang@kaust.edu.sa

***email:* ying.sun@kaust.edu.sa

SUMMARY: When performing spatial regression analysis in environmental data applications, spatial heterogeneity in the regression coefficients is often observed. Spatially varying coefficient models, including geographically weighted regression and spline models, are standard tools for quantifying such heterogeneity. In this paper, we propose a spatially varying coefficient model that represents the spatially varying parameters as a mixture of local polynomials at selected locations. The local polynomial parameters have attractive interpretations, indicating various types of spatial heterogeneity. Instead of estimating the spatially varying regression coefficients directly, we develop a penalized least squares regression procedure for the local polynomial parameter estimation, which both shrinks the parameter estimation and penalizes the differences among parameters that are associated with neighboring locations. We develop confidence intervals for the varying regression coefficients and prediction intervals for the response. We apply the proposed method to characterize the spatially varying association between particulate matter concentrations ($PM_{2.5}$) and pollutant gases related to the secondary aerosol formulation in China. The identified regression coefficients show distinct spatial patterns for nitrogen dioxide, sulfur dioxide, and carbon monoxide during different seasons.

KEY WORDS: Fine particular matter; Fused Lasso; Spatial regression; Spatially varying coefficient model.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi:10.1111/biom.13077

This article is protected by copyright. All rights reserved.

1. Introduction

Particulate matter, especially fine particulate matter with an aerodynamic diameter less than $2.5\ \mu\text{m}$ ($\text{PM}_{2.5}$), is an air pollutant that poses great risk to human health (World Health Organization, 2003). Long-term exposure to heavy $\text{PM}_{2.5}$ pollution is positively associated with lung cancer, and cardiovascular problems (Hu et al., 2008), while short-term exposure is associated with respiratory illnesses, and total hospital admissions (Lu et al., 2015). Understanding the drivers of $\text{PM}_{2.5}$ is critical for developing pollution control strategies and public health policies. In this paper, we analyze a dataset of air pollution levels in China collected by the Center for Geographic Analysis, Harvard University. We are interested in modeling the relationship between $\text{PM}_{2.5}$ and the level of air pollutants such as the nitrogen dioxide (NO_2) and the sulfur dioxide (SO_2), which are precursors of $\text{PM}_{2.5}$ (Huang et al., 2014), and investigating the spatial patterns of the contributions of these pollutants.

The relationships among these environmental variables have been widely studied in the literature, such as the relationship between $\text{PM}_{2.5}$ and meteorological variables (Russell et al., 2017) and the relationship between $\text{PM}_{2.5}$ and NO_2 , SO_2 (Huang et al., 2014). Usually, these relationships vary across the studied region. Fotheringham et al. (2002) introduced the geographically weighted regression (GWR) to account for spatial heterogeneity in the regression coefficients. In this paper, we aim to develop better tools for analyzing spatial heterogeneity in regression coefficients for environmental data.

A well-known tool for handling coefficient heterogeneity is the varying coefficient model. In the statistics literature, varying coefficient models in which the coefficient functions vary with a single variable, such as time, are well studied. See Fan and Zhang (2008) and Hastie and Tibshirani (1993) for the method based on the kernel smoother and the spline smoother, respectively. However, these results have not been extended to geographical regression models.

Bayesian analysis is an important class of methods for spatial modeling, see Carlin et al. (2014) for an introduction. Gelfand et al. (2003, 2005) developed a Bayesian hierarchical framework for spatial point-reference data, for which they proposed to model the varying regression coefficients as Gaussian processes. Assunção (2003) studied a Bayesian varying coefficient model (BVCM) for areal data. Wheeler et al. (2014) compared BVCM with GWR, they found that BVCM has better estimation and prediction accuracy in some applications. More recently, Kim and Lee (2017) generalized BVCM to handle geographically hierarchical data. Franco-Villoria et al. (2018) used penalized complexity priors to properly shrink BVCM to a simpler model such that overfitting can be avoided.

The most commonly used GWR for spatial data is essentially a locally constant kernel smoother (Fan and Gijbels, 1996). However, GWR has several drawbacks. For example, the estimated regression coefficients suffer from the boundary effects: the bias is of a higher order at the boundary than in the interior of the studied region (Fan and Gijbels, 1996). Also, GWR tends to produce ragged and biased estimates (Wang et al., 2008). In addition, a single bandwidth as used in GWR is not sufficient to accommodate the curvature of multiple spatially varying coefficients (Leong and Yue, 2017).

To address these drawbacks of GWR, Wang et al. (2008) proposed a local linear GWR which reduced the bias of the coefficient estimates compared to basic GWR. Leong and Yue (2017) proposed a conditional GWR, in which they selected a different bandwidth for each coefficient function. Russell et al. (2017) proposed a penalized local linear estimator that selected covariates at each spatial location of interest.

In addition to the drawbacks mentioned above, GWR and its extensions have further limitations. First, they cannot explore local homogeneity in the regression coefficients. For example, if the regression coefficient is constant in the neighborhood of a spatial location, they cannot detect the constancy of the regression coefficient. Second, GWR and its extensions are implemented in a location-by-location manner, i.e., the regression coefficients have to be re-estimated at each spatial location of interest. For example, if a 1000×1000 grid of spatial locations is of interest to draw a fine map of the regression coefficients, then we must implement the estimation process 10^6 times.

Motivated by these observations, we propose a spatially varying coefficient model that represents the spatially varying parameters as a mixture of local polynomials at selected locations, which we refer to as anchor points. Instead of estimating the spatially varying regression coefficients directly, we develop a penalized least squares regression procedure that is more flexible and more efficient than GWR. We penalize the differences among parameters that are associated with neighboring anchor points using the Lasso (Tibshirani, 1996) and fused Lasso (Tibshirani et al., 2005) penalties to explore local homogeneity.

The proposed methodology overcomes the drawbacks of GWR and its extensions. We consider local polynomial approximations rather than local constant approximations used in GWR. The penalty imposed on the local polynomial coefficients allows us to borrow information from neighboring spatial locations, and accommodates both homogeneity and heterogeneity in the regression coefficients. Thus, our method is more efficient and more flexible in recovering the underlying functional relationship. For example, if the coefficient function is constant over all or part of the region, our method

can recover the constant relationship automatically. Also, the regression function can be interpolated at any location of interest; no estimation is involved in this step. Therefore, the computational burden of estimating the coefficient function at a large number of spatial locations is much relieved compared to GWR and especially penalized GWR.

The remainder of the paper is organized as follows. In Section 2, we introduce the spatial penalized local polynomial regression model and investigate its theoretical properties. In Section 3, we develop an algorithm and propose a criterion for selecting the tuning parameters. We examine the performance of the proposed methodology in a simulation study in Section 4, and present a PM_{2.5} data analysis in Section 5. Section 6 concludes the paper.

2. Methodology

2.1 Spatial penalized local polynomial regression model

Suppose we observe $(y(\mathbf{s}_i), \mathbf{x}(\mathbf{s}_i)^\top), i = 1, \dots, n$ at locations $\mathbf{s}_i = (s_{i1}, s_{i2}) \in D \subset \mathbb{R}^2$, where $y(\mathbf{s}_i)$ is the response variable, and $\mathbf{x}(\mathbf{s}_i) \in \mathbb{R}^q$ is a vector of covariates. A spatially varying coefficient regression model has the form

$$y(\mathbf{s}) = \mathbf{x}(\mathbf{s})^\top \boldsymbol{\beta}(\mathbf{s}) + \epsilon(\mathbf{s}), \text{ for } \mathbf{s} \in D, \quad (1)$$

where the coefficients $\boldsymbol{\beta}(\mathbf{s})$ vary spatially. To simplify the notation, we proceed by assuming that there is only one covariate in model (1). That is, $\beta(\mathbf{s})$ is one-dimensional. The general case can be treated similarly with more complicated notation.

To estimate the spatially varying regression coefficients $\beta(\mathbf{s})$, we introduce a set of anchor points, $\mathbf{s}_1^*, \dots, \mathbf{s}_m^* \in D$, where $\mathbf{s}_i^* = (s_{i1}^*, s_{i2}^*)$. Without further information, we suggest that the anchor points should be placed regularly across the region D under study. We approximate the regression coefficient $\beta(\mathbf{s})$ at an arbitrary geographical location $\mathbf{s} = (s_1, s_2) \in D$ by a weighted average of the Taylor expansion up to order p at each anchor point. Specifically, we approximate $\beta(\mathbf{s})$ at each $\mathbf{s}_i^*$ as $\beta_i(\mathbf{s}) \approx \sum_{v_1+v_2 \leq p} \theta_{iv_1v_2} (s_1 - s_{i1}^*)^{v_1} (s_2 - s_{i2}^*)^{v_2}$. Further, we approximate $\beta(\mathbf{s})$ as a weighted average of $\beta_i(\mathbf{s})$:

$$\beta(\mathbf{s}) \approx \sum_{i=1}^m w_i(\mathbf{s}) \beta_i(\mathbf{s}). \quad (2)$$

The weights $w_i(\mathbf{s}), i = 1, \dots, m$, are defined as $w_i(\mathbf{s}) = K_{h_n}(\|\mathbf{s} - \mathbf{s}_i^*\|) / \{\sum_{j=1}^m K_{h_n}(\|\mathbf{s} - \mathbf{s}_j^*\|)\}$, where $K_{h_n}(\cdot)$ is a kernel function (we use the Gaussian kernel in this work), h_n is a bandwidth, and $\|\cdot\|$ is an appropriate distance in $\mathbb{R}^2$. Because the approximation is a linear combination of the parameters $\boldsymbol{\theta} = (\theta_{iv_1v_2}, i = 1 \dots m, v_1 + v_2 \leq p)$, we write it as $x(\mathbf{s})\beta(\mathbf{s}) := \mathbf{z}(\mathbf{s})^\top \boldsymbol{\theta}$, where $\mathbf{z}(\mathbf{s})$ is the pseudo covariates. The idea of (2) is that when $\mathbf{s}$ is

close to one anchor point $\mathbf{s}_i$, the Taylor expansion at $\mathbf{s}_i$ should be more representative of $\beta(\mathbf{s})$. The approximation in (2) relies on all the anchor points such that the information in the Taylor expansion at all the anchor points, especially those close to $\mathbf{s}$, can be leveraged to better approximate $\beta(\mathbf{s})$.

Let $\mathbf{y} = (y(\mathbf{s}_1), \dots, y(\mathbf{s}_n))^\top$ and $\mathbf{Z} = (\mathbf{z}_1^\top, \dots, \mathbf{z}_n^\top)^\top$. With approximation (2), model (1) is transformed into a linear model with the parameter $\boldsymbol{\theta}$. The transformed model can be estimated by the ordinary least squares regression. However, we estimate the parameter $\boldsymbol{\theta}$ by the penalized least squares regression, such that the spatial homogeneity among the elements of $\boldsymbol{\theta}$ can be utilized. We propose the spatial penalized local polynomial regression estimator as follows:

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\theta}\|_2^2 + \sum_{v_1+v_2 \leq p} \sum_{\|\mathbf{s}_i^* - \mathbf{s}_j^*\| \leq d} \frac{\lambda}{\|\mathbf{s}_i^* - \mathbf{s}_j^*\|^\gamma} |\theta_{iv_1v_2} - \theta_{jv_1v_2}| \right\}, \quad (3)$$

where λ is a tuning parameter, and $\|\cdot\|_2$ is the Euclidean norm. The penalty function is adapted from the fused LASSO (Tibshirani et al., 2005). We note that the parameter $\boldsymbol{\theta}_i = (\theta_{iv_1v_2}, v_1 + v_2 \leq p)$ in approximation (2) is the coefficient of the Taylor expansion with respect to the i th anchor point $\mathbf{s}_i^*$. For anchor points that are close to each other, we expect that their corresponding coefficients $\boldsymbol{\theta}_i$ do not differ much from each other. We penalize $|\theta_{iv_1v_2} - \theta_{jv_1v_2}|$ to recover the spatial homogeneity among the elements of $\boldsymbol{\theta}$. The penalty can also be weighted by the distance between the anchor points such that penalties on neighboring coefficients with a shorter distance get more weight.

To fix ideas, in the rest of the paper, we consider the case where the threshold d is selected such that only the differences of coefficients from adjacent anchor points are penalized, and we consider the simple case where the weights are all equal, which is equivalent to set $\gamma = 0$. Furthermore, we write the sum of the absolute differences $\sum_{v_1+v_2 \leq p} \sum_{\|\mathbf{s}_i^* - \mathbf{s}_j^*\| \leq d} (|\theta_{iv_1v_2} - \theta_{jv_1v_2}|)$ in a concise form $\|\mathbf{A}\boldsymbol{\theta}\|_1$, where the matrix $\mathbf{A}$ represents the contrasts and $\|\cdot\|_1$ is the l_1 norm. We introduce a LASSO penalty (Tibshirani, 1996) to remove insignificant coefficients. The objective function can be simplified to be

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\theta}\|_2^2 + \lambda \|\mathbf{A}\boldsymbol{\theta}\|_1 + \mu \|\boldsymbol{\theta}\|_1 \right\}. \quad (4)$$

We do not distinguish the estimators (3) and (4), because the estimator (3) is a special case of (4) with $\mu = 0$. We include 0 as a candidate value for μ , and select the parameters λ and μ using the BIC criterion, which is detailed in Section 3.

2.2 Confidence interval and prediction interval

In this section, we develop confidence intervals for the regression coefficients and prediction intervals for the response. Inference methods for penalized regression models is a recent and less developed research topic. The first line of research focuses on providing confidence intervals for low-dimensional parameters of the model, e.g., a single regression coefficient. This line of research includes multi sample-splitting (Meinshausen et al., 2009), regularized projection (Zhang and Zhang, 2014), and residual bootstrap methods (Chatterjee et al., 2013). Another line of research focuses on testing linear contrasts of the mean of the response variable conditional on a selected model (Lee and Taylor, 2014; Tibshirani et al., 2018). See Dezeure et al. (2015) for a comprehensive review.

We adapt Efron (2014)'s method to our setting and propose confidence intervals for the regression coefficient $\beta(\mathbf{s})$ at an arbitrary location $\mathbf{s}$. Efron (2014) used the bagging estimator to smooth discontinuities due to model selection, and he developed a standard deviation formula for the bagging estimator. We describe the procedure briefly as follows. We draw B non-parametric bootstrap samples from the observed data. Let $F_i^*, i = 1, \dots, B$ be vectors of length n recording the number of times of the original sample occurring in the bootstrap samples. We obtain the bootstrap replications $\hat{\beta}(\mathbf{s})_i^*, i = 1, \dots, B$ based on the bootstrap samples. Define the bagging estimator as $\tilde{\beta}(\mathbf{s}) = \sum_{i=1}^B \hat{\beta}(\mathbf{s})_i^* / B$, Efron (2014)'s formula for the standard deviation of $\tilde{\beta}(\mathbf{s})$ is $\tilde{\text{sd}}(\tilde{\beta}(\mathbf{s})) = \left(\sum_{j=1}^n \widehat{\text{cov}}_j^2 \right)^{1/2}$, where $\widehat{\text{cov}}_j = \sum_{i=1}^B (\hat{\beta}(\mathbf{s})_i^* - \tilde{\beta}(\mathbf{s})) (F_{ij}^* - F_{.j}^*) / B$ and $F_{.j}^* = \sum_{i=1}^B F_{ij}^* / B$. The confidence interval for $\beta(\mathbf{s})$, named smoothed interval, with the coverage level $1 - \alpha$ can be constructed as $\tilde{\beta}(\mathbf{s}) \pm z_{1-\alpha/2} \tilde{\text{sd}}(\tilde{\beta}(\mathbf{s}))$, where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

The conformal prediction developed in Vovk et al. (2005) is a general framework for prediction inference in regression analysis. It can generate distribution-free prediction intervals using any estimator of the regression function with finite sample coverage guarantees even when the model is misspecified (Lei et al., 2018). The conformal interval is close to an optimal interval under weak conditions (Lei and Wasserman, 2014). We use the split conformal prediction (Lei et al., 2018), a low-computational variant, to construct prediction intervals for our model. Suppose we have a new observation $x(\mathbf{s})$, to construct a prediction interval for $y(\mathbf{s})$ we proceed as follows. Randomly split the observed data into two equally sized subsets, fit the model using one subset, and obtain absolute residuals using the other. Let d be the $1 - \alpha$ sample quantile of the absolute residuals, a level $1 - \alpha$ prediction interval for $y(\mathbf{s})$ is $x(\mathbf{s})\hat{\beta}(\mathbf{s}) \pm d$, where $\hat{\beta}(\mathbf{s})$ is obtained using only the fitting sample.

2.3 Properties of the proposed model

In this section, we discuss the theoretical properties of the proposed model. We investigate the consistency and the convergence rate of the proposed estimator under the assumption that approximation (2) is exact. That is, we assume that the regression coefficient $\beta(\mathbf{s})$ can be represented exactly as in (2). We focus on the estimation error of $\hat{\boldsymbol{\theta}}$. When approximation error exists, we investigate the performance of $\hat{\boldsymbol{\theta}}$ numerically in Section 4.

First, we formally describe the concept of homogeneity of the elements of $\boldsymbol{\theta}$. Let r denote the length of $\boldsymbol{\theta}$. We assume that the set $\{1, 2, \dots, r\}$ is partitioned into K subsets $\mathcal{A} = \{A_1, \dots, A_K\}$, and $|A_k|$ denotes the size of the set A_k . Let $\mathcal{M}_A = \{\boldsymbol{\theta} \in \mathbb{R}^r : \theta_i = \theta_j, \text{ if } i, j \in A_k, k = 1, \dots, K\}$ denote a subspace of $\mathbb{R}^r$. By homogeneity, we mean $\boldsymbol{\theta} \in \mathcal{M}_A$, where A_k is the index set in which all θ_j share the same value when the indices $j \in A_k$. The next theorem states the consistency and convergence results for the estimator $\hat{\boldsymbol{\theta}}$ in (3).

THEOREM 1: *Under the regularity conditions listed in Ke et al. (2015), the estimator $\hat{\boldsymbol{\theta}}$ has the following properties:*

- (1) $\mathbb{P}(\hat{\boldsymbol{\theta}} \in \mathcal{M}_A) \rightarrow 1$ as the sample size $n \rightarrow \infty$;
- (2) $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| = O_p(\sqrt{K/n} + \lambda(\sum_{k=1}^K 1/|A_k|)^{1/2})$.

The first part of Theorem 1 guarantees that the estimator $\hat{\boldsymbol{\theta}}$ can recover the homogeneity structure of $\boldsymbol{\theta}$ asymptotically. The second part of Theorem 1 states the convergence rate of the estimator $\hat{\boldsymbol{\theta}}$. As pointed out by Ke et al. (2015), the term $\lambda(\sum_{k=1}^K 1/|A_k|)^{1/2}$ is of the order $\sqrt{K \log(n \vee r)/n}$. Thus, the estimator $\hat{\boldsymbol{\theta}}$ is consistent as long as $K \log(r)$ is asymptotically negligible compared to the sample size n . A least squares estimator has the convergence rate $\sqrt{r/n}$. Therefore, the proposed estimator converges faster than the least squares estimator since $K \leq r$. The gain in convergence rate is more significant for a smaller K ; that is, when $\boldsymbol{\theta}$ can be partitioned into a small number of homogeneous groups.

3. Computational Issues

3.1 Estimation algorithm

In this section, we propose an algorithm for minimizing the objective function (4). First, we introduce a set of nuisance parameters, $\boldsymbol{\eta}_1 = \mathbf{A}\boldsymbol{\theta}$ and $\boldsymbol{\eta}_2 = \boldsymbol{\theta}$, and rewrite the

This article is protected by copyright. All rights reserved.

objective function (4) as a constrained minimization problem,

$$Q(\boldsymbol{\theta}, \boldsymbol{\eta}) = \frac{1}{2} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\eta}_1\|_1 + \mu \|\boldsymbol{\eta}_2\|_1,$$

subject to $\mathbf{A}\boldsymbol{\theta} - \boldsymbol{\eta}_1 = 0, \boldsymbol{\theta} - \boldsymbol{\eta}_2 = 0,$

where $\boldsymbol{\eta} = (\boldsymbol{\eta}_1^\top, \boldsymbol{\eta}_2^\top)^\top$. To solve the constrained minimization problem, we implement the alternating direction method of multipliers (ADMM) algorithm (Boyd et al., 2011). The augmented Lagrangian is written as

$$L(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{v}) = Q(\boldsymbol{\theta}, \boldsymbol{\eta}) + \rho \mathbf{v}_1^\top (\mathbf{A}\boldsymbol{\theta} - \boldsymbol{\eta}_1) + \frac{\rho}{2} \|\mathbf{A}\boldsymbol{\theta} - \boldsymbol{\eta}_1\|_2^2 + \rho \mathbf{v}_2^\top (\boldsymbol{\theta} - \boldsymbol{\eta}_2) + \frac{\rho}{2} \|\boldsymbol{\theta} - \boldsymbol{\eta}_2\|_2^2,$$

where $\mathbf{v} = (\mathbf{v}_1^\top, \mathbf{v}_2^\top)^\top$ are the Lagrangian multipliers and ρ is a penalty parameter in the ADMM algorithm. Let $(\boldsymbol{\theta}^{(1)}, \boldsymbol{\eta}^{(1)}, \mathbf{v}^{(1)})$ be initial values of the parameters. We update the parameters as follows until convergence:

$$\begin{aligned} \boldsymbol{\theta}^{(k+1)} &= \arg \min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{\eta}^{(k)}, \mathbf{v}^{(k)}), \\ \boldsymbol{\eta}^{(k+1)} &= \arg \min_{\boldsymbol{\eta}} L(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\eta}, \mathbf{v}^{(k)}), \\ \mathbf{v}_1^{(k+1)} &= \mathbf{v}_1^{(k)} + \mathbf{A}\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\eta}_1^{(k+1)}, \quad \mathbf{v}_2^{(k+1)} = \mathbf{v}_2^{(k)} + \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\eta}_2^{(k+1)}. \end{aligned}$$

To update $\boldsymbol{\theta}$, we need to solve the least squares minimization problem

$$\boldsymbol{\theta}^{(k+1)} = \arg \min_{\boldsymbol{\theta}} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\theta}\|_2^2 + \frac{\rho}{2} \|\mathbf{A}\boldsymbol{\theta} - \boldsymbol{\eta}_1^{(k)} + \mathbf{v}_1^{(k)}\|_2^2 + \frac{\rho}{2} \|\boldsymbol{\theta} - \boldsymbol{\eta}_2^{(k)} + \mathbf{v}_2^{(k)}\|_2^2 \right\}.$$

It can be solved efficiently by QR decomposition, and the decomposition can only be done once and used repeatedly in the iteration.

To update $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$, we need to solve the minimization problems

$$\begin{aligned} \boldsymbol{\eta}_1^{(k+1)} &= \frac{\rho}{2} \|\mathbf{A}\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\eta}_1 + \mathbf{v}_1^{(k)}\| + \lambda \|\boldsymbol{\eta}_1\|_1, \text{ and} \\ \boldsymbol{\eta}_2^{(k+1)} &= \frac{\rho}{2} \|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\eta}_2 + \mathbf{v}_2^{(k)}\| + \mu \|\boldsymbol{\eta}_2\|_1. \end{aligned}$$

These minimization problems have explicit solutions (Fan and Li, 2001). Define the soft thresholding function as $S(x; \lambda) = (|x| - \lambda)_+ \text{sign}(x)$, where $(x)_+$ is the positive part of x and $\text{sign}(x)$ is the sign of x . When the argument x is a vector, the thresholding function is implemented in an element-wise way. Then, $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ can be updated as

$$\boldsymbol{\eta}_1^{(k+1)} = S(\mathbf{A}\boldsymbol{\theta}^{(k+1)} + \mathbf{v}_1^{(k)}; \lambda), \text{ and } \boldsymbol{\eta}_2^{(k+1)} = S(\boldsymbol{\theta}^{(k+1)} + \mathbf{v}_2^{(k)}; \mu).$$

The convergence of the ADMM algorithm can be monitored by the primal and dual residuals, which are defined as $r^{(k)} = \|\mathbf{A}\boldsymbol{\theta}^{(k)} - \boldsymbol{\eta}_1^{(k)}\| + \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\eta}_2^{(k)}\|$, and $s^{(k)} = \|\mathbf{v}^{(k)} - \mathbf{v}^{(k-1)}\|$, respectively. The algorithm is terminated when the primal residual and the dual residual

are both less than a threshold, for example, 10^{-6} . The linear convergence property of the ADMM algorithm is well established in the literature (Nishihara et al., 2015).

3.2 Choice of the tuning parameter

To implement the proposed method, we need to select the bandwidth parameter h , and the tuning parameters λ and μ . We propose selecting these parameters by minimizing the modified Bayesian Information Criterion (mBIC) as in Wang et al. (2007),

$$mBIC(h, \lambda, \mu) = \log \left\{ \frac{1}{n} \left\| \mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\theta}} \right\|_2^2 \right\} + 0.5 \frac{\log n}{n} \hat{N}, \quad (5)$$

where the parameter estimate $\hat{\boldsymbol{\theta}}$ depends on the parameters h , λ , and μ , and the parameter $\hat{N}$ is the number of unique elements in $\hat{\boldsymbol{\theta}}$. For example, if $\hat{\boldsymbol{\theta}} = (0.1, 0.1, 0.2, 0.2)$, then $\hat{N} = 2$. The mBIC criterion balances the model fit and the parsimony of the parameters. In practice, we minimize the mBIC criterion on a regular grid of the tuning parameter values.

4. Simulation Study

4.1 Simulation with approximation error

In this section, we illustrate how the proposed method approximates a smooth function. Four smooth regression coefficients, denoted as cases (a)-(d), defined on $[0, 1] \times [0, 1]$ are considered, see details in Web Appendix A.1. These functions represent a variety of behaviors of smooth functions, e.g., different number of peaks and valleys, roughness, and range.

Here, we illustrate the approximation error produced when reconstructing a smooth function using the proposed method. To achieve this, we sample $n = 200$ and 500 points uniformly from the unit square $[0, 1] \times [0, 1]$ and generate the data from $y = \beta(s_1, s_2)$. No noise is included. The test functions are reconstructed using the proposed method for different orders of the Taylor expansion ($p = 0, 1, 2$) and numbers of anchor points ($m = 9, 16, 25, 36, 49$). We summarize the results below, and present the detailed results in Web Table 1. The root mean square error (RMSE) is defined as $RMSE = \sqrt{\frac{1}{10201} \sum_{i=1}^{10201} (\beta(\mathbf{s}_i) - \hat{\beta}(\mathbf{s}_i))^2}$, where the sum is taken over a 101×101 regular grid in the unit square $[0, 1] \times [0, 1]$. The RMSE decreases as both the number of anchor points and the order of Taylor expansion increase. Increasing the order of the Taylor expansion from $p = 0$ to $p = 1$ significantly reduces RMSE, whereas increasing from $p = 1$ to $p = 2$ does not further reduce RMSE, except in the cases (a) and (c) for a small number of anchor points ($m \leq 16$). The effect of increasing the number of anchor points vanishes

This article is protected by copyright. All rights reserved.

gradually. Based on these observations, we recommend two strategies for fitting real data: (1) use $p = 0$ and a large number of anchor points (more than 49); (2) use $p = 1$ and a moderate number of anchor points (36 or 49).

Next, we compare the proposed method with GWR and the local linear GWR (LL) (Wang et al., 2008). We also compare with the Bayesian varying coefficient model (BVCM) of Franco-Villoria et al. (2018) implemented by the R package INLA (Martins et al., 2013) where varying regression coefficients are modeled as Gaussian processes with Matérn covariance functions and penalized complexity priors. For the proposed method, we use $m = 36$ anchor points which are placed on a regular grid. We refer to the case where the order of Taylor expansion is $p = 0$ as WLC (weighted locally constant), and $p = 1$ as WLL (weighted local linear). The data are generated from the model $y_i = 1 + x_i\beta(\mathbf{s}_i) + \epsilon_i$, $i = 1, \dots, n$, where the covariate x is generated from the standard normal distribution $N(0, 1)$ and the error is independently drawn from $N(0, \sigma^2)$, where $\sigma = 0.1$ and $\sigma = 0.5$. The coefficient function $\beta(\mathbf{s})$ takes the form of each of the four cases (a) through (d) as in the last example. The locations $\mathbf{s}_i, i = 1, \dots, n$ are sampled uniformly from $[0, 1] \times [0, 1]$. The sample size is $n = 200$ and $n = 500$ and the simulation is repeated 100 times.

We see that RMSEs for WLC and WLL are much lower than either GWR or LL for all the test functions and for all combinations of the sample size n and the error standard deviation σ (Table 1). For example, when $\beta(s)$ takes the form of (c) and $\sigma = 0.1$, RMSEs for the sample size $n = 200$ are 0.225 and 0.176 for WLC and WLL, respectively, and reduce to 0.157 and 0.098 when the sample size increases to $n = 500$. In contrast, RMSEs are much larger, 0.627 and 0.415 for GWR and LL, respectively, for the sample size $n = 200$, and 0.367 and 0.187 for the sample size $n = 500$. The main reason that GWR does not work as well as our method is that it does not balance bias and variance well. From Web Tables 3 and 4, we observe that GWR tends to have a higher bias than its variance. Also the bias of GWR is higher than that of the proposed method. BVCM is comparable to, and in case (d) better than, the proposed method in terms of RMSE.

Next, we compare the proposed method with GWR and LL in terms of prediction error. In the simulation, an independent test sample of size 10000 is drawn from the true model. Predictions are made using the estimated coefficients, and the root mean squared prediction error (RMSPE) is defined as $\text{RMSPE} = \sqrt{\frac{1}{10000} \sum_{i=1}^{10000} (y(\mathbf{s}_i) - \hat{y}(\mathbf{s}_i))^2}$. In summary, RMSPEs are much lower for WLC and WLL than for GWR and LL in all

cases (Figure 1). The linear approximations have a lower prediction error. For example, for the test function (c), RMSPE of WLL is smaller than that of WLC, and RMSPE of LL is smaller than that of the locally constant estimator in GWR.

We constructed 95% confidence intervals for $\beta(\mathbf{s})$ on a 101×101 regular grid with $B = 200$ bootstrap samples and 95% prediction intervals for the response for a new sample of size 10000. Detailed results are shown in Web Appendix Section A.1. The empirical coverage probabilities for $\beta(\mathbf{s})$ are close or greater than the nominal level 95% except for case (c) when the estimator is WLC. Confidence intervals based on WLL are better than those based on WLC in terms of coverage probabilities, especially for the case (c) and the locations close to peaks, valleys and boundaries. The empirical coverage probabilities of the prediction intervals are all close to or greater than the nominal level 95% for all cases. For example, the empirical coverage probabilities based on WLL are 97.5%, 95.9%, 97.1%, and 95.7% for cases (a)-(d) respectively when $n = 500$ and $\sigma = 0.1$.

4.2 Simulation without approximation error

In this section, the regression coefficients are defined as $\beta(\mathbf{s}) = \sum_{i=1}^m w_i \{\theta_{i0} + \theta_{i1}(s_1 - s_{i1}^*) + \theta_{i2}(s_2 - s_{i2}^*)\}$, and the rest of the data generating process is the same as the last section. Three different cases of θ , denoted as (a), (b), and (c), are considered, where details are shown in Web Appendix Section A.2. The coefficients θ in cases (a), (b), and (c) have different degrees of sparsity and similarity, implying that the coefficient function $\beta(\mathbf{s})$ has different degrees of roughness and homogeneity.

We estimate the regression coefficient β by WLC and WLL with 36 regularly spaced anchor points. For the purpose of comparison, we estimate β using the proposed method with the true anchor points, and the resulting estimator is denoted by PLS (penalized least squares). The model is also estimated without any penalty on parameter θ , and the resulting estimator is denoted by LS (least squares). We also compare the proposed estimator with GWR, LL, and BVCM as in the last section.

Comparing PLS with LS, PLS has a lower RMSE for each regression coefficient β for the case (a) in which the true coefficients θ are both sparse and locally homogeneous (Table 2). For the other cases, PLS has a lower RMSE for the intercept. We conclude that the proposed penalized method significantly reduces the estimation error compared to LS when the true coefficient is sparse or locally similar.

Comparing WLC and WLL with GWR and LL, RMSE is significantly higher for GWR and LL than for the proposed methods, WLC and WLL, in all cases. On average, RMSE

This article is protected by copyright. All rights reserved.

is three times larger for GWR than for WLC or WLL. WLC and WLL are also better than BVCM in terms of RMSE. When estimating the constant intercept, RMSE is lower for WLC and WLL than for LS, which assumes knowledge of the true anchor points. That is because the proposed penalized method automatically recovers the constant intercept, and the estimation error is significantly reduced.

Next, we compare all the methods in terms of the prediction error. RMSPEs of WLC and WLL are comparable to those of PLS and LS which assume knowledge of the true anchor points (Figure 2). RMSPE is much higher for GWR and LL than for the proposed methods in all cases. RMSPE is also higher for BVCM than for the proposed methods in all cases.

The empirical coverage probabilities of the confidence intervals for $\beta(\mathbf{s})$ are greater than the nominal level 95% for all the cases (Web Figures 10, 11, and 12). The empirical coverage probabilities of prediction intervals based on WLL are 95.0%, 95.7%, 96.0% for case (a), (b), and (c) respectively when $n = 500$ and $\sigma = 0.1$. The empirical coverage probabilities and the length of prediction intervals based on WLC do not differ from significantly from those based on WLL.

5. Analysis of the Air Pollution Data

Secondary aerosol formulation is an important source of $\text{PM}_{2.5}$; it occurs when primary gases such as sulfur (SO_2) and nitrogen oxides (NO_2) turn into sulfuric acid and nitric acid. Huang et al. (2014) found that severe haze pollution events in China were driven largely by secondary aerosol formation, which contributed 30–77% of the total $\text{PM}_{2.5}$, and also that the contribution of secondary aerosols was different in north China and south China.

We analyze the spatially varying relationship between the level of $\text{PM}_{2.5}$ and the level of the pollutants NO_2 , SO_2 , and CO using the proposed method. The data are available at the website <http://aqicn.org/>. The data were collected at more than 2000 monitoring locations between January 2014 and January 2016. The study region in our analysis corresponds to the eastern and central China. To show the seasonal differences among the drivers of $\text{PM}_{2.5}$, we consider two time periods, January 2015 and July 2015, with 1139 and 918 observation locations, respectively. The data are averaged at each location within each time period.

This article is protected by copyright. All rights reserved.

We model the relationship between $\text{PM}_{2.5}$ and NO_2 , SO_2 , and CO as

$$\text{PM}_{2.5}(\mathbf{s}) = \beta_0(\mathbf{s}) + \text{NO}_2(\mathbf{s})\beta_1(\mathbf{s}) + \text{SO}_2(\mathbf{s})\beta_2(\mathbf{s}) + \text{CO}(\mathbf{s})\beta_3(\mathbf{s}) + \epsilon(\mathbf{s}). \quad (6)$$

We estimate the spatially varying coefficients $\boldsymbol{\beta}(\mathbf{s}) = (\beta_0(\mathbf{s}), \beta_1(\mathbf{s}), \beta_2(\mathbf{s}), \beta_3(\mathbf{s}))^\top$ using the proposed method with local linear expansion and 80 anchor points placed on a regular grid. The locations of the anchor points and the observations are shown in Web Appendix B Figure 13. The tuning parameters are selected according to the mBIC criterion (5), we used the same shrinkage parameters for all the covariates. Besides the proposed method, we also estimate $\boldsymbol{\beta}(\mathbf{s})$ using GWR for comparison. The 95% confidence intervals for the coefficients are shown in Web Figures 14 and 15. The purpose of this study is to show the spatially varying relationship between $\text{PM}_{2.5}$ and the pollutants NO_2 , SO_2 , and CO in different seasons. The estimated spatially varying coefficients are shown in Figure 3 and Figure 4.

In Figures 3 and 4, we see that the estimated coefficients $\boldsymbol{\beta}(\mathbf{s})$ vary spatially throughout different seasons. The intercept, which can be interpreted as the baseline pollution level, is estimated to be a constant $103.97 \mu\text{g}/\text{m}^3$ in January 2015 (Figure 3(a)). The NO_2 level has a high correspondence with $\text{PM}_{2.5}$ in the middle of China, i.e., Chongqing, Hubei, and Hunan provinces, as shown in Figure 3(a). The SO_2 level has a high correspondence with $\text{PM}_{2.5}$ in Shanghai and Jiangsu province. The CO level has a high correspondence with $\text{PM}_{2.5}$ in Beijing and Hebei province. Compared to GWR (Figure 3(b)), the coefficients estimated by the proposed method are more smooth with clearly identified spatial patterns; the coefficients estimated by GWR are rough and noisy across the studied region. Therefore, interesting spatial patterns are difficult to identify in the GWR-estimated coefficients.

For July 2015, the intercept is estimated to be a constant $62.49 \mu\text{g}/\text{m}^3$, which reflects a lower baseline $\text{PM}_{2.5}$ level than in January 2015. The NO_2 and SO_2 coefficients are significant in Beijing and Hebei province (Figure 4(a)). The CO coefficient is constant across the studied region, except for the Guangdong and Fujian Provinces (Figure 4(a)). The coefficients estimated by GWR reflect hardly any useful information (Figure 4(b)). To compare the accuracy of prediction made by the proposed method and GWR, we conduct a 10-fold cross-validation by randomly splitting the data into ten subsets. We calculate point estimates and 95% prediction intervals of the response for the left-

This article is protected by copyright. All rights reserved.

out data. For January 2015, RMSPE is 21.64 for the proposed method and 23.18 for GWR. For July 2015, RMSPE is 18.22 for the proposed method and 19.11 for GWR. The proposed method has a lower prediction error than GWR. The empirical coverage probabilities of the prediction intervals for the left-out samples are 95.7% and 97.1% for the data in January 2015 and July 2015 respectively.

6. Conclusion

In this work, we propose a penalized local polynomial model for the estimation of spatially varying coefficient models. Based on the weighted local polynomial expansion on a grid of anchor points, we penalize the differences among coefficients in the expansion to recover the possible local homogeneity structure for the regression coefficients. We develop confidence intervals for the varying regression coefficients and prediction intervals for the response. Compared to GWR, the proposed estimator has high estimation and prediction accuracy.

In the $PM_{2.5}$ data analysis, the locations of the pollution data were not uniformly distributed, there were more observations in areas close to the coastline, and comparatively few inland observations were available. The proposed method does not utilize the information in the density of the spatial locations. Extending the current methodology to accommodate observations that are not uniformly distributed can improve its performance in real data analysis. Furthermore, extending the theoretical results to cases in which the approximation errors exist for an unknown smooth surface requires future research. We did not consider models with correlated errors, for which the estimation and theoretical studies are much more challenging.

ACKNOWLEDGEMENTS

We thank the associate editor and a referee for their help in improving the paper. This research was supported by funding from King Abdullah University of Science and Technology (KAUST).

REFERENCES

Assunção, R. M. (2003). Space varying coefficient models for small area data. *Environmetrics* **14**, 453–473.

This article is protected by copyright. All rights reserved.

- Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* **3**, 1–122.
- Carlin, B. P., Gelfand, A. E., and Banerjee, S. (2014). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.
- Chatterjee, A., Lahiri, S., et al. (2013). Rates of convergence of the adaptive LASSO estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics* **41**, 1232–1259.
- Dezeure, R., Bühlmann, P., Meier, L., Meinshausen, N., et al. (2015). High-dimensional inference: Confidence intervals, p-values and R-software hdi. *Statistical science* **30**, 533–558.
- Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association* **109**, 991–1007.
- Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.
- Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and its Interface* **1**, 179–195.
- Fotheringham, A. S., Brunson, C., and Charlton, M. (2002). *Geographically weighted regression: the analysis of spatially varying relationships*. New York: Wiley, USA.
- Franco-Villoria, M., Ventrucci, M., and Rue, H. (2018). Bayesian varying coefficient models using PC priors. *arXiv preprint arXiv:1806.02084*.
- Gelfand, A. E., Banerjee, S., and Gamerman, D. (2005). Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics* **16**, 465–479.
- Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association* **98**, 387–396.
- Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B* **55**, 757–796.
- Hu, Z., Liebens, J., and Rao, K. R. (2008). Linking stroke mortality with air pollution, income, and greenness in northwest Florida: an ecological geographical study. *International journal of health geographics* **7**, 20.
- Huang, R.-J., Zhang, Y., Bozzetti, C., Ho, K.-F., Cao, J.-J., Han, Y., et al. (2014). High secondary aerosol contribution to particulate pollution during haze events in China.

- Nature* **514**, 218–222.
- Ke, Z. T., Fan, J., and Wu, Y. (2015). Homogeneity pursuit. *Journal of the American Statistical Association* **110**, 175–194.
- Kim, H. and Lee, J. (2017). Hierarchical spatially varying coefficient process model. *Technometrics* **59**, 521–527.
- Lee, J. and Taylor, J. (2014). Exact post model selection inference for marginal screening. *Advances in Neural Information Processing Systems* **27**, 136–144.
- Lei, J., G’Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association* **113**, 1094–1111.
- Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B* **76**, 71–96.
- Leong, Y.-Y. and Yue, J. C. (2017). A modification to geographically weighted regression. *International journal of health geographics* **16**, 11.
- Lu, F., Xu, D., Cheng, Y., Dong, S., Guo, C., Jiang, X., and Zheng, X. (2015). Systematic review and meta-analysis of the adverse health effects of ambient PM_{2.5} and PM₁₀ pollution in the Chinese population. *Environmental research* **136**, 196–204.
- Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). Bayesian computing with inla: new features. *Computational Statistics & Data Analysis* **67**, 68–83.
- Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association* **104**, 1671–1681.
- Nishihara, R., Lessard, L., Recht, B., Packard, A., and Jordan, M. I. (2015). A general analysis of the convergence of admm. *arXiv preprint arXiv:1502.02009*.
- Russell, B. T., Wang, D., and McMahan, C. S. (2017). Spatially modeling the effects of meteorological drivers of PM_{2.5} in the Eastern United States via a local linear penalized quantile regression estimator. *Environmetrics* **28**, e2448.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* **58**, 267–288.
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society. Series B* **67**, 91–108.
- Tibshirani, R. J., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2018). Uniform asymptotic inference and the bootstrap after model selection. *The Annals of Statistics* **46**, 1255–1287.
- Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random*

- World*. Springer.
- Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568.
- Wang, N., Mei, C.-L., and Yan, X.-D. (2008). Local linear estimation of spatially varying coefficient models: an improvement on the geographically weighted regression technique. *Environment and Planning A* **40**, 986–1005.
- Wheeler, D. C., Páez, A., Spinney, J., and Waller, L. A. (2014). A bayesian approach to hedonic price analysis. *Papers in Regional Science* **93**, 663–683.
- World Health Organization (2003). Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide. Report on a WHO working group, Bonn, Germany.
- Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B* **76**, 217–242.

SUPPORTING INFORMATION

Web Appendix A and B, referenced in Section 4 and 5, are available with this paper at the Biometrics website on Wiley Online Library. The R package for the proposed methodology is available as an online supporting information.

Figure 1. RMSPE for the case where $\sigma = 0.1$ for the simulation with approximation error in Section 4.1. This figure appears in color in the electronic version of this article.

Figure 2. RMSPE for the case where $\sigma = 0.1$ for the simulation without approximation error in Section 4.2. This figure appears in color in the electronic version of this article.

Figure 3. The estimated spatially varying coefficients for modeling $PM_{2.5}$ in January 2015. Column (a) shows the results obtained by the proposed method, and column (b) shows those obtained by GWR. This figure appears in color in the electronic version of this article.

Figure 4. The estimated spatially varying coefficients for modeling $PM_{2.5}$ in July 2015. Column (a) shows the results obtained by the proposed method, and column (b) shows those obtained by GWR. This figure appears in color in the electronic version of this article.

Table 1*RMSE when $\sigma = 0.1$ for the simulation with approximation error; standard errors are shown in parentheses.*

	$\beta(\mathbf{s})$	WLC	WLL	GWR	LL	BVCM
n=200	(a)	0.106(0.018)	0.109(0.020)	0.332(0.028)	0.242(0.023)	0.107(0.015)
	(b)	0.056(0.010)	0.053(0.010)	0.125(0.015)	0.098(0.012)	0.057(0.008)
	(c)	0.225(0.048)	0.175(0.044)	0.627(0.074)	0.415(0.065)	0.177(0.038)
	(d)	0.057(0.010)	0.057(0.011)	0.074(0.011)	0.075(0.009)	0.046(0.005)
n=500	(a)	0.063(0.007)	0.061(0.008)	0.206(0.012)	0.123(0.007)	0.065(0.006)
	(b)	0.036(0.006)	0.032(0.005)	0.082(0.007)	0.061(0.006)	0.037(0.004)
	(c)	0.157(0.016)	0.098(0.015)	0.367(0.037)	0.187(0.017)	0.093(0.014)
	(d)	0.040(0.004)	0.040(0.004)	0.051(0.005)	0.049(0.005)	0.032(0.004)

Table 2

RMSE ($\times 10^{-2}$) when $\sigma = 0.1$ for the simulation without approximation error; standard errors ($\times 10^{-2}$) are shown in parentheses.

		n=200			n=500		
		(a)	(b)	(c)	(a)	(b)	(c)
PLS	β_1	0.80 (0.60)	1.98(0.69)	1.85(0.68)	0.52(0.38)	1.06(0.36)	1.13(0.36)
	β_2	1.26 (0.58)	2.97(0.89)	2.96(1.00)	0.76(0.38)	1.64(0.40)	1.68(0.52)
	β_3	1.30 (0.60)	2.96(0.75)	2.97(0.80)	0.70(0.40)	1.70(0.43)	1.75(0.51)
LS	β_1	2.73 (0.57)	2.85(0.64)	2.72(0.61)	1.57(0.36)	1.60(0.34)	1.61(0.35)
	β_2	2.88 (0.64)	2.97(0.82)	3.06(0.72)	1.72(0.37)	1.63(0.37)	1.59(0.38)
	β_3	3.04 (0.85)	2.99(0.81)	3.08(0.78)	1.63(0.35)	1.72(0.41)	1.67(0.40)
WLC	β_1	0.73 (0.51)	1.76(0.69)	2.27(0.67)	0.47(0.33)	1.06(0.39)	1.28(0.40)
	β_2	1.79 (0.53)	3.74(0.84)	3.93(1.08)	1.25(0.28)	2.57(0.45)	2.31(0.45)
	β_3	1.78 (0.62)	2.78(0.53)	4.00(0.96)	1.23(0.31)	1.93(0.38)	2.44(0.51)
WLL	β_1	0.77 (0.54)	1.72(0.68)	2.26(0.71)	0.47(0.34)	1.01(0.40)	1.31(0.37)
	β_2	1.75 (0.47)	3.72(0.80)	3.92(0.99)	1.23(0.33)	2.54(0.40)	2.28(0.42)
	β_3	1.77 (0.61)	2.82(0.56)	4.02(1.01)	1.17(0.30)	1.95(0.37)	2.41(0.51)
GWR	β_1	5.88 (1.23)	6.53(1.05)	7.35(1.13)	3.56(0.49)	4.10(0.52)	4.79(0.49)
	β_2	9.35 (1.63)	13.74(1.90)	15.24(2.10)	5.67(0.65)	8.33(1.07)	9.40(0.91)
	β_3	9.27 (1.38)	9.04(1.32)	14.79(2.00)	5.69(0.72)	5.54(0.58)	9.11(0.88)
LL	β_1	3.55 (0.61)	3.61(0.66)	3.98(0.70)	2.24(0.29)	2.31(0.32)	2.53(0.31)
	β_2	6.91 (0.87)	8.56(1.00)	7.40(1.07)	4.24(0.44)	5.19(0.53)	4.38(0.53)
	β_3	6.84 (0.86)	5.08(0.86)	9.04(1.12)	4.21(0.41)	3.21(0.47)	5.22(0.53)
BVCM	β_1	1.14 (0.62)	1.07(0.50)	1.14(0.55)	0.60(0.34)	0.68(0.31)	0.65(0.34)
	β_2	4.06 (0.63)	4.89(0.77)	5.42(0.87)	2.70(0.35)	3.22(0.37)	3.37(0.40)
	β_3	4.04 (0.59)	3.82(0.71)	5.67(0.86)	2.72(0.34)	2.53(0.31)	3.57(0.40)