# Model Selection and Structure Specification in Ultra-High Dimensional Generalised Semi-Varying Coefficient Models

**Article** *in* The Annals of Statistics · June 2015

DOI: 10.1214/15-AOS1356

CITATIONS
17

READS
238

**3 authors**, including:

Degui Li
The University of York
**69** PUBLICATIONS **465** CITATIONS

SEE PROFILE

Yuan Ke
University of Georgia
**14** PUBLICATIONS **74** CITATIONS

SEE PROFILE

*Submitted to the Annals of Statistics*

# MODEL SELECTION AND STRUCTURE SPECIFICATION IN ULTRA-HIGH DIMENSIONAL GENERALISED SEMI-VARYING COEFFICIENT MODELS*

By Degui Li , Yuan Ke and Wenyang Zhang†

*University of York, UK*

In this paper, we study the model selection and structure specification for the generalised semi-varying coefficient models (GSVCMs), where the number of potential covariates is allowed to be larger than the sample size. We first propose a penalised likelihood method with the LASSO penalty function to obtain the preliminary estimates of the functional coefficients. Then, using the quadratic approximation for the local log-likelihood function and the adaptive group LASSO penalty (or the local linear approximation of the group SCAD penalty) with the help of the preliminary estimation of the functional coefficients, we introduce a novel penalised weighted least squares procedure to select the significant covariates and identify the constant coefficients among the coefficients of the selected covariates, which could thus specify the semiparametric modelling structure. The developed model selection and structure specification approach not only inherits many nice statistical properties from the local maximum likelihood estimation and non-concave penalised likelihood method, but also computationally attractive thanks to the computational algorithm that is proposed to implement our method. Under some mild conditions, we establish the asymptotic properties for the proposed model selection and estimation procedure such as the sparsity and oracle property. We also conduct simulation studies to examine the finite sample performance of the proposed method, and finally apply the method to analyse a real data set, which leads to some interesting findings.

**1. Introduction.** In recent years, model selection has become an important and fundamental issue in data analysis as high-dimensional data are commonly encountered in various applied fields such as epidemiology, genetics and finance. It is well-known that the traditional model selection

*This research was supported by the Singapore National Research Foundation under its Cooperative Basic Research Grant and administered by the Singapore Ministry of Health National Medical Research Council (Grant No. NMRC/CBRG/0014/2012) and the National Science Foundation of China (Grant No. 11271242)

†Correspondent author. Email: wenyang.zhang@york.ac.uk

*AMS 2000 subject classifications:* Primary 62G08; secondary 62G20

*Keywords and phrases:* GSVCM, LASSO, local maximum likelihood, oracle estimation, SCAD, sparsity, ultra-high dimension

procedures such as the stepwise regression and the best subset variable selection can be extremely computationally intensive in the analysis of the high-dimensional data. To address this computational challenge, various penalised likelihood/least-square methods have been well studied and become a promising alternative. With an appropriate penalty function, the penalised method would automatically shrink the small coefficients to zero and remove the associated variables from the model, hence serve the purpose of model selection. Some commonly-used penalty functions include the LASSO (Tibshirani 1996), SCAD (Fan and Li, 2001), group LASSO (Yuan and Lin, 2006), adaptive LASSO (Zou, 2006) and MCP (Zhang, 2010), and the algorithms to implement the penalised likelihood/least squares methods have also been developed in the literature (c.f., Efron *et al*, 2004; Hunter and Li, 2005; Zou and Li, 2008). In high-dimensional data analysis, it is often the case that the number of potential covariates grows over sample size or even diverges with certain exponential rate. In the context of parametric models, there has been some literature addressing this problem, see, for example, Huang and Xie (2007), Fan and Lv (2008), Huang *et al* (2008), Zhang and Huang (2008), Fan *et al* (2009), Zou and Zhang (2009), Fan and Song (2010) and Bühlmann and van de Geer (2011).

However, the pre-supposed parametric linear relationships and models, although easy to implement, are often too restricted and unrealistic in practical application. They often lead to model misspecification, which would result in inconsistent estimates and incorrect conclusions being drawn from the data analysed. In this paper, we relax this linear restriction and use functional coefficients to describe the relationship between response and covariates. The varying coefficient models, as an important and useful generalisation of the linear models, have played a very important role in the analysis of the complex data and experienced deep and exciting developments, see, for example, Fan and Zhang (1999, 2000), Cheng *et al* (2009), Wang and Xia (2009), Wang *et al* (2009), Zhang *et al* (2009), Kai *et al* (2011) and Li and Zhang (2011). Suppose we have a response variable $y$, an index variable $U$, and potential covariates $X = (x_1, \cdots, x_{d_n})^{\mathrm{T}}$, where the dimension $d_n$ depends on sample size $n$ and $d_n \to \infty$ when $n \to \infty$. Define the conditional expectation of $y$ for given $(U, X)$ by

$$m(U, X) = \mathsf{E}(y|U, X).$$

We assume in this paper that the log conditional density function of $y$ given

$X$ and $U$ is

$$C_1(\boldsymbol{\phi}_1)\ell\big(m(U,X),y\big)+C_2(y,\boldsymbol{\phi}_2)\ \text{ with }\ g\big(m(U,\ X)\big)=\sum_{j=1}^{d_n}a_j(U)x_j, \tag{1.1}$$

where $g(\cdot)$, $\ell(\cdot,\ \cdot)$, $C_1(\cdot)$ and $C_2(\cdot,\ \cdot)$ are known, the functional coefficients $a_1(\cdot),\cdots,a_p(\cdot)$ are unknown to be estimated, $C_1(\boldsymbol{\phi}_1)>0$, $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ are unknown nuisance parameters. When the response variable is discrete, we define the density function as its probability mass function. It is easy to see that model (1.1) is a natural extension of the generalised linear models by allowing the coefficients varying with the index variable $U$. As some functional coefficients in (1.1) may be constant coefficients, we call (1.1) as generalised semi-varying coefficient models (GSVCMs).

The model selection in the varying coefficient models (which can be seen as a special case of the GSVCMs) has been extensively studied in existing literature. For instance, Wang *et al* (2008) and Wang and Xia (2009) use the group penalisation to select the significant variables in the varying coefficient models when the number of potential covariates is fixed. More recently, for the ultra-high dimensional varying coefficient models, Song *et al* (2012), Cheng *et al* (2014), Fan *et al* (2014) and Liu *et al* (2014) combine the nonparametric independence screening technique and the group penalised method to choose the significant covariates and estimate the functional coefficients for the varying coefficient models. Fan *et al* (2011) consider a nonparametric independence screening in sparse ultra-high dimensional additive models. Wei *et al* (2011) consider the penalised variable selection by using a basis function approximation for the functional coefficients in the varying coefficient models and allows that the number of covariates diverges with the sample size. Lian (2012) further generalises Wei *et al* (2011)'s methodology to the generalised varying coefficient models which are similar to our framework (1.1). Unlike the existing literature, in this paper, the model selection for the proposed GSVCMs has two aspects: (1) variable selection; and (2) identification of the constant coefficients. As the variable selection is equivalent to identifying the zero functional coefficients, and the identification of the constant coefficients is equivalent to identifying the functional coefficients with zero derivative or variation. Either of the two aspects would be related to the so-called "all-in-all-out" problem.

In this paper, we first propose a penalised likelihood method with the LASSO penalty function to obtain the preliminary estimates of the functional coefficients, which is proved to be uniformly consistent. The uniform convergence rate for the preliminary penalised nonparametric estima-

tion results relies on the number of non-zero functional coefficients and the tuning parameter involved in the penalty term. Then, we use the preliminary estimates of the functional coefficients in the quadratic approximation for the local log-likelihood function and the construction of the adaptive group LASSO penalty or the local linear approximation of the group SCAD penalty, and introduce a novel penalised weighted least squares procedure to simultaneously select the significant covariates and identify the constant coefficients among the coefficients of the selected covariates. Hence, the semi-varying coefficient modelling structure can be specified. The developed model selection and structure specification approach inherits many nice statistical properties from both the local maximum likelihood estimation and non-concave penalised likelihood method. Under some regularity conditions, we establish the asymptotic properties for the proposed model selection and estimation procedure such as the sparsity and oracle property. In order to implement our method in practice, we develop a novel computational algorithm to do the maximisation involved in the estimation procedure when the SCAD or LASSO penalty is used. The SCAD has many advantages, and is widely used as a penalty function in the shrinkage methods. The commonly-used approach, to deal with the SCAD penalty in the implementation of shrinkage method for the varying coefficient models, consists of two steps: (1) approximate SCAD by an $L_1$ penalty locally by local linear approximation; (2) apply the quadratic approximation to deal with the $L_1$ penalty. In this paper, we do not go down that route. Making use of the structure of the group SCAD, we propose a different algorithm to implement our method. Our simulation results show that both the adaptive group LASSO and the SCAD methods perform reasonably well with the latter giving slightly better performance, and the method developed in the present paper outperforms those in Wang and Xia (2009), and Lian (2012).

The rest of the paper is organised as follows. Section 2 describes the penalised model selection and structure specification procedure. Section 3 gives the asymptotic properties of the proposed model selection and structure specification procedure. Section 4 provides a computational algorithm to implement the developed method and discusses how to determine the tuning parameters. Section 5 compares the finite sample performance of the developed model selection with those proposed in the existing literature through some simulation studies. In Section 6, we apply the GSVCMs together with the proposed model selection, structure specification and estimation procedure to analyse an environmental data set from Hong Kong, and explore how some pollutants and other environmental factors affect the number of daily total hospital admissions for circulationary and respiration-

ary problems in Hong Kong. The regularity conditions for the asymptotic theory are given in Appendix A. The proofs of the main theoretical results and some auxiliary results are provided in Appendices B and C of a supplemental document [Li *et al* (2015)].

**2. Model selection and structure specification method.** For any function $f(\cdot)$, throughout this paper, we use $\dot{f}(\cdot)$ to denote its first-order derivative, and $\ddot{f}(\cdot)$ its second-order derivative. For any vector $\mathbf{u}$, we define $\|\mathbf{u}\|^2 = \mathbf{u}^{\mathrm{T}}\mathbf{u}$. As in the generalised linear models, our main interest lies in the conditional mean of the response variable for given covariates, and $C_1(\boldsymbol{\phi}_1)$ and $C_2(y, \boldsymbol{\phi}_2)$ in model (1.1) have little to do with the mean part. In order to make the presentation simpler, without loss of generality, we assume the log conditional density function of $y$ given $X$ and $U$ is

$$\ell\big(m(U,X),y\big) \quad \text{with} \quad g\big(m(U,X)\big) = \sum_{j=1}^{d_n} a_j(U)x_j \tag{2.1}$$

and further assume the support of the index variable $U$ is $[0,1]$ throughout this paper. Our model selection and structure specification procedure can be summarised as follows: (i) use the penalised local maximum likelihood method with the LASSO penalty to obtain the preliminary estimation of the functional coefficients (see Section 2.1); (ii) consider the quadratic approximation of the log-likelihood function by using the preliminary functional coefficients estimates and the approximated log-likelihood function is essentially an $L_2$ objective function (see Section 2.2); (iii) conduct the variable selection and structure specification by using a penalised weighted least squares method with two types of weighted group LASSO penalty functions: the adaptive group LASSO and the local linear approximation of the group SCAD where the preliminary functional coefficients estimates are also used (see Section 2.3); (iv) finally estimate the constant coefficients in the GSVCMs (see Section 2.4). The model selection procedure proposed in this paper can be seen, in some sense, as a generalisation of Fan *et al* (2014)'s folded concave penalised estimation for ultra-high dimensional parametric regression models.

2.1. *Preliminary estimation of the functional coefficients.* Suppose we have a sample $(U_i, X_i, y_i)$, $i = 1, \cdots, n$, from model (2.1), where $X_i = (x_{i1}, \cdots, x_{id_n})^{\mathrm{T}}$. For each given $k$, $k = 1, \cdots, n$, by Taylor's expansion of $a_j(\cdot)$, $j = 1, \cdots, d_n$, we have

$$a_j(U_i) \approx a_j(U_k) + \dot{a}_j(U_k)(U_i - U_k),$$

when $U_i$, $i = 1, \cdots, n$, are in a small neighbourhood of $U_k$. This local linear approximation leads to the construction of the following local log-likelihood function to estimate $a_j(U_k)$ and $\dot{a}_j(U_k)$, $j = 1, \cdots, d_n$,

(2.2)
$$\mathcal{L}_{nk}(\mathbf{a}_k, \mathbf{b}_k) = \frac{1}{n}\sum_{i=1}^{n} \ell\Big(g^{-1}\Big\{\sum_{j=1}^{d_n}\big[\alpha_{jk} + \beta_{jk}(U_i - U_k)\big]x_{ij}\Big\},\ y_i\Big)K_h(U_i - U_k),$$

where $K(\cdot)$ is a kernel function, $h$ is a bandwidth, $K_h(\cdot) = \frac{1}{h}K(\cdot/h)$,

$$\mathbf{a}_k = \big(\alpha_{1k},\ \cdots,\ \alpha_{d_nk}\big)^{\mathrm{T}}, \quad \mathbf{b}_k = \big(\beta_{1k},\ \cdots,\ \beta_{d_nk}\big)^{\mathrm{T}}.$$

When the dimension of the covariates is fixed, we may obtain the solution which maximises the local log-likelihood function $\mathcal{L}_{nk}(\cdot,\cdot)$ defined in (2.2) and show that the resulting nonparametric estimators are consistent (c.f., Cai *et al*, 2000; Zhang and Peng, 2010). However, for the case of the ultra-high dimensional GSVCMs, it would be difficult to obtain satisfactory estimation by maximising $\mathcal{L}_{nk}(\cdot,\cdot)$ as the number of unknown nonparametric components involved exceeds the number of observations. In order to address this issue, we next introduce a penalised local log-likelihood method by adding an appropriate penalty function to the above local log-likelihood function.

Without loss of generality, we assume that there exist $1 \le s_{n1} < s_{n2} < d_n$ such that for $1 \le j \le s_{n1}$, $a_j(\cdot)$ are the functional coefficients with non-zero deviation; for $s_{n1} + 1 \le j \le s_{n2}$, $a_j(\cdot) \equiv c_j$ are the constant coefficients; for $s_{n2} + 1 \le j \le d_n$, $a_j(\cdot) \equiv 0$. Moreover, we assume that $s_{n2}$, although may diverge with the sample size, is much smaller than the sample size $n$ and the dimension of the whole covariates $d_n$. Hence, for any $k = 1, \cdots, n$, the number of non-zero elements in $\mathbf{a}_{k0} = \big[a_1(U_k),\ \cdots,\ a_{d_n}(U_k)\big]^{\mathrm{T}}$ and $\mathbf{b}_{k0} = \big[\dot{a}_1(U_k),\ \cdots,\ \dot{a}_{d_n}(U_k)\big]^{\mathrm{T}}$ is at most $s_{n1} + s_{n2}$. Define the penalised local log-likelihood function with the LASSO penalty function as

$$\mathcal{Q}_{nk}(\mathbf{a}_k, \mathbf{b}_k) = \mathcal{L}_{nk}(\mathbf{a}_k, \mathbf{b}_k) - \lambda_1\sum_{j=1}^{d_n}|\alpha_{jk}| - \lambda_2\sum_{j=1}^{d_n}h|\beta_{jk}|, \tag{2.3}$$

where $\lambda_1$ and $\lambda_2$ are two tuning parameters. We let $(\widetilde{\mathbf{a}}_k,\ \widetilde{\mathbf{b}}_k)$ be the maximiser of $\mathcal{Q}_{nk}(\cdot,\cdot)$ and call it the preliminary estimator of the functional coefficients $a_j(U_k)$s and their derivatives $\dot{a}_j(U_k)$s, $j = 1, \cdots, d_n$.

We will show in Proposition 3.1 that the above preliminary estimator obtained by the penalised local likelihood estimation with the LASSO penalty

is uniformly consistent. The preliminary estimates of the functional coefficients will be used in the approximation of log-likelihood function and the construction of weighted LASSO penalty functions in our model selection and structure specification procedure, see Sections 2.2 and 2.3 below.

2.2. *Quadratic approximation of the log-likelihood estimation.* In the model selection and structure specification procedure for the GSVCMs, we need to consider the following local log-likelihood function:

$$\mathcal{L}_n(\mathcal{A},\ \mathcal{B}) = \sum_{k=1}^{n} \mathcal{L}_{nk}(\mathbf{a}_k, \mathbf{b}_k) \tag{2.4}$$

where $\mathcal{A} = \big(\mathbf{a}_1^{\mathrm{T}},\ \cdots,\ \mathbf{a}_n^{\mathrm{T}}\big)^{\mathrm{T}}$, $\mathcal{B} = \big(\mathbf{b}_1^{\mathrm{T}},\ \cdots,\ \mathbf{b}_n^{\mathrm{T}}\big)^{\mathrm{T}}$, and $\mathcal{L}_{nk}(\mathbf{a}_k, \mathbf{b}_k)$ is defined in (2.2). To alleviate the computational burden for the optimisation of $\mathcal{L}_n(\mathcal{A},\ \mathcal{B})$, we next introduce a simple approximation.

Let

$$\dot{\mathcal{L}}_n(\mathcal{A},\ \mathcal{B}) = \Big[\dot{\mathcal{L}}_{n1}^{\mathrm{T}}(\mathbf{a}_1, \mathbf{b}_1), \cdots, \dot{\mathcal{L}}_{nn}^{\mathrm{T}}(\mathbf{a}_n, \mathbf{b}_n)\Big]^{\mathrm{T}}$$

and

$$\ddot{\mathcal{L}}_n(\mathcal{A},\ \mathcal{B}) = \mathsf{diag}\left\{\ddot{\mathcal{L}}_{n1}(\mathbf{a}_1, \mathbf{b}_1), \cdots, \ddot{\mathcal{L}}_{nn}(\mathbf{a}_n, \mathbf{b}_n)\right\},$$

where

$$\begin{aligned}
\dot{\mathcal{L}}_{nk}(\mathbf{a}_k, \mathbf{b}_k) &= \frac{1}{n}\sum_{i=1}^{n} q_1\Big\{\sum_{j=1}^{d_n}\big[\alpha_{jk} + \beta_{jk}(U_i - U_k)\big]x_{ij},\ y_i\Big\}\begin{pmatrix} X_i \\ \frac{U_i - U_k}{h}\cdot X_i \end{pmatrix}\cdot \\
&\quad K_h(U_i - U_k), \\
\ddot{\mathcal{L}}_{nk}(\mathbf{a}_k, \mathbf{b}_k, l) &= \frac{1}{n}\sum_{i=1}^{n} q_2\Big\{\sum_{j=1}^{d_n}\big[\alpha_{jk} + \beta_{jk}(U_i - U_k)\big]x_{ij},\ y_i\Big\}\left(\frac{U_i - U_k}{h}\right)^{l}\cdot \\
&\quad X_i X_i^{\mathrm{T}} K_h(U_i - U_k), \quad l = 0, 1, 2, \\
\ddot{\mathcal{L}}_{nk}(\mathbf{a}_k, \mathbf{b}_k) &= \begin{bmatrix} \ddot{\mathcal{L}}_{nk}(\mathbf{a}_k, \mathbf{b}_k, 0) & \ddot{\mathcal{L}}_{nk}(\mathbf{a}_k, \mathbf{b}_k, 1) \\ \ddot{\mathcal{L}}_{nk}(\mathbf{a}_k, \mathbf{b}_k, 1) & \ddot{\mathcal{L}}_{nk}(\mathbf{a}_k, \mathbf{b}_k, 2) \end{bmatrix},
\end{aligned}$$

and

$$q_1(s, y) = \frac{\partial \ell\big[g^{-1}(s), y\big]}{\partial s}, \quad q_2(s, y) = \frac{\partial^2 \ell\big[g^{-1}(s), y\big]}{\partial s^2}.$$

Denote $\widetilde{\mathcal{A}}_n = \big(\widetilde{\mathbf{a}}_1^{\mathrm{T}}, \cdots, \widetilde{\mathbf{a}}_n^{\mathrm{T}}\big)^{\mathrm{T}}$ and $\widetilde{\mathcal{B}}_n = \big(\widetilde{\mathbf{b}}_1^{\mathrm{T}}, \cdots, \widetilde{\mathbf{b}}_n^{\mathrm{T}}\big)^{\mathrm{T}}$ where $(\widetilde{\mathbf{a}}_k, \widetilde{\mathbf{b}}_k)$ is the maximiser of the objective function $\mathcal{Q}_{nk}(\cdot, \cdot)$ in (2.3), and define

$$\mathcal{V}_n(\mathcal{A}, \mathcal{B}) = \big(\mathbf{a}_1^{\mathrm{T}}, \mathbf{b}_1^{\mathrm{T}}, \cdots, \mathbf{a}_n^{\mathrm{T}}, \mathbf{b}_n^{\mathrm{T}}\big)^{\mathrm{T}}, \quad \mathcal{V}_n(\mathcal{A}, h\mathcal{B}) = \big(\mathbf{a}_1^{\mathrm{T}}, h\mathbf{b}_1^{\mathrm{T}}, \cdots, \mathbf{a}_n^{\mathrm{T}}, h\mathbf{b}_n^{\mathrm{T}}\big)^{\mathrm{T}}.$$

By Taylor's expansion of the log-likelihood function $\mathcal{L}_n(\mathcal{A},\ \mathcal{B})$, we may obtain the following quadratic approximation:

$$\begin{aligned}\mathcal{L}_n(\mathcal{A},\ \mathcal{B}) &\approx \mathcal{L}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n) + \big[\mathcal{V}_n(\mathcal{A}, h\mathcal{B}) - \mathcal{V}_n(\widetilde{\mathcal{A}}_n, h\widetilde{\mathcal{B}}_n)\big]^{\mathrm{T}} \dot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n) + \\ &\quad \frac{1}{2}\big[\mathcal{V}_n(\mathcal{A}, h\mathcal{B}) - \mathcal{V}_n(\widetilde{\mathcal{A}}_n, h\widetilde{\mathcal{B}}_n)\big]^{\mathrm{T}} \ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n)\big[\mathcal{V}_n(\mathcal{A}, h\mathcal{B}) - \mathcal{V}_n(\widetilde{\mathcal{A}}_n, h\widetilde{\mathcal{B}}_n)\big] \\ &\equiv \mathcal{L}_n^{\diamond}(\mathcal{A},\ \mathcal{B}). \qquad (2.5)\end{aligned}$$

It is easy to see that $\mathcal{L}_n^{\diamond}(\mathcal{A},\ \mathcal{B})$ is essentially an $L_2$ objective function. Hence, it would be much easier to deal with $\mathcal{L}_n^{\diamond}(\mathcal{A},\ \mathcal{B})$ in (2.5) than to directly deal with $\mathcal{L}_n(\mathcal{A},\ \mathcal{B})$. In the model selection procedure introduced in Section 2.3 below, we may replace $\mathcal{L}_n(\mathcal{A},\ \mathcal{B})$ by $\mathcal{L}_n^{\diamond}(\mathcal{A},\ \mathcal{B})$.

2.3. *Penalised local likelihood estimation with weighted LASSO penalties.* In order to conduct the model selection and structure specification, we define the following penalised local log-likelihood function:

$$\mathcal{Q}_n(\mathcal{A},\ \mathcal{B}) = \mathcal{L}_n(\mathcal{A},\ \mathcal{B}) - \sum_{j=1}^{d_n} p_{nj}\big(\|\boldsymbol{\alpha}_j\|\big) - \sum_{j=1}^{d_n} p_{nj}^{*}\big(\|\boldsymbol{\beta}_j\|\big), \qquad (2.6)$$

where $p_{nj}(\cdot)$ and $p_{nj}^{*}(\cdot)$ are two penalty functions which will be specified later,

$$\boldsymbol{\alpha}_j = (\alpha_{j1},\ \cdots,\ \alpha_{jn})^{\mathrm{T}} \quad \text{and} \quad \boldsymbol{\beta}_j = (\beta_{j1},\ \cdots,\ \beta_{jn})^{\mathrm{T}},$$

which correspond to $[a_j(U_1),\ \cdots,\ a_j(U_n)]^{\mathrm{T}}$ and $[\dot{a}_j(U_1),\ \cdots,\ \dot{a}_j(U_n)]^{\mathrm{T}}$, respectively. By the quadratic approximation (2.5), we may approximate $\mathcal{Q}_n(\mathcal{A},\ \mathcal{B})$ by $\mathcal{Q}_n^{\diamond}(\mathcal{A},\ \mathcal{B})$ which is defined through

$$\mathcal{Q}_n^{\diamond}(\mathcal{A},\ \mathcal{B}) = \mathcal{L}_n^{\diamond}(\mathcal{A},\ \mathcal{B}) - \sum_{j=1}^{d_n} p_{nj}\big(\|\boldsymbol{\alpha}_j\|\big) - \sum_{j=1}^{d_n} p_{nj}^{*}\big(\|\boldsymbol{\beta}_j\|\big), \qquad (2.7)$$

where $\mathcal{L}_n^{\diamond}(\mathcal{A},\ \mathcal{B})$ is defined in (2.5).

For the penalty functions $p_{nj}(\cdot)$ and $p_{nj}^{*}(\cdot)$ in (2.6) and (2.7), we consider two possible cases: (i) the adaptive group LASSO penalty, and (ii) the group SCAD penalty. Note that identifying the constant coefficients in model (2.1) is equivalent to identifying the functional coefficients such that either $\dot{a}_j(U_1) = \cdots = \dot{a}_j(U_n) = 0$ or its deviation $D_j = 0$, where

$$D_j = \Big\{ \sum_{k=1}^{n} \big[a_j(U_k) - \frac{1}{n}\sum_{l=1}^{n} a_j(U_l)\big]^2 \Big\}^{1/2}.$$

Using the preliminary estimates, we can construct the preliminary estimator of $D_j$:

$$\widetilde{D}_j = \Big\{ \sum_{k=1}^{n} \big[\widetilde{a}_j(U_k) - \frac{1}{n}\sum_{l=1}^{n} \widetilde{a}_j(U_l)\big]^2 \Big\}^{1/2},$$

where $\widetilde{a}_j(U_k)$ is the $j$-th element of $\widetilde{\mathbf{a}}_k$.

For case (i) of the adaptive group LASSO, we define

$$p_{nj}\big(\|\boldsymbol{\alpha}_j\|\big) = \lambda_3 \|\widetilde{\boldsymbol{\alpha}}_j\|^{-\kappa} \|\boldsymbol{\alpha}_j\|, \quad p^*_{nj}\big(\|\boldsymbol{\beta}_j\|\big) = \lambda_3^* \widetilde{D}_j^{-\kappa} \|h\boldsymbol{\beta}_j\|, \tag{2.8}$$

where $\lambda_3$ and $\lambda_3^*$ are two tuning parameters, $\kappa$ is pre-determined and can be chosen as 1 or 2 as in the literature, $\widetilde{\boldsymbol{\alpha}}_j = [\widetilde{a}_j(U_1), \cdots, \widetilde{a}_j(U_n)]^{\mathrm{T}}$.

For case (ii) of the group SCAD, we may apply the local linear approximation to the SCAD penalty function $p_{nj}(\cdot)$ (Zou and Li, 2008) and then obtain

$$p_{nj}\left(\|\boldsymbol{\alpha}_j\|\right) \approx p_{nj}\left(\|\widetilde{\boldsymbol{\alpha}}_j\|\right) - \dot{p}_{nj}\left(\|\widetilde{\boldsymbol{\alpha}}_j\|\right)\|\widetilde{\boldsymbol{\alpha}}_j\| + \dot{p}_{nj}\left(\|\widetilde{\boldsymbol{\alpha}}_j\|\right)\|\boldsymbol{\alpha}_j\|, \tag{2.9}$$

where $p_{nj}(z) \equiv p_{\lambda_4}(z)$ is the SCAD penalty function with the derivative defined by

$$\dot{p}_{nj}(z) \equiv \dot{p}_{\lambda_4}(z) = \lambda_4\big[I(z \le \lambda_4) + \frac{(a_0\lambda_4 - z)_+}{(a_0-1)\lambda} I(z > \lambda_4)\big], \tag{2.10}$$

$\lambda_4$ is a tuning parameter and $a_0 = 3.7$ as suggested in Fan and Li (2001). Note that the first two terms on the right hand side of (2.9) do not involve $\|\boldsymbol{\alpha}_j\|$, which motivates us to choose

$$p_{nj}\left(\|\boldsymbol{\alpha}_j\|\right) = \dot{p}_{\lambda_4}\left(\|\widetilde{\boldsymbol{\alpha}}_j\|\right)\|\boldsymbol{\alpha}_j\| \tag{2.11}$$

with $\dot{p}_{\lambda_4}(\cdot)$ defined in (2.10). For $p^*_{nj}\big(\|\boldsymbol{\beta}_j\|\big)$, similar to the corresponding definition in (2.8) for case (i), we consider the structure:

$$p^*_{nj}\big(\|\boldsymbol{\beta}_j\|\big) = \dot{p}_{\lambda_4^*}(\widetilde{D}_j)\|h\boldsymbol{\beta}_j\|, \tag{2.12}$$

where $\dot{p}_{\lambda_4^*}(\cdot)$ is defined similarly to $\dot{p}_{\lambda_4}(\cdot)$ in (2.10) with $\lambda_4$ replaced by $\lambda_4^*$.

Based on (2.7) and the above specification of the penalty functions, we may consider the following two objective functions:

$$\mathcal{Q}_n^1(\mathcal{A},\ \mathcal{B}) = \mathcal{L}_n^{\diamond}(\mathcal{A},\ \mathcal{B}) - \lambda_3 \sum_{j=1}^{d_n} \|\widetilde{\boldsymbol{\alpha}}_j\|^{-\kappa}\|\boldsymbol{\alpha}_j\| - \lambda_3^* \sum_{j=1}^{d_n} \widetilde{D}_j^{-\kappa}\|h\boldsymbol{\beta}_j\| \tag{2.13}$$

for the adaptive group LASSO penalty; and

$$\mathcal{Q}_n^2(\mathcal{A},\ \mathcal{B}) = \mathcal{L}_n^{\diamond}(\mathcal{A},\ \mathcal{B}) - \sum_{j=1}^{d_n} \dot{p}_{\lambda_4}(\|\widetilde{\boldsymbol{\alpha}}_j\|)\|\boldsymbol{\alpha}_j\| - \sum_{j=1}^{d_n} \dot{p}_{\lambda_4^*}(\widetilde{D}_j)\|h\boldsymbol{\beta}_j\| \tag{2.14}$$

for the group SCAD penalty. Note that the penalty terms in (2.13) and (2.14) are the weighted LASSO penalty functions. In particular, the weights in (2.14) are determined by the derivative of the SCAD penalty using the preliminary estimators $\widetilde{\boldsymbol{\alpha}}_j$ and $\widetilde{D}_j$. The objective functions in (2.13) and (2.14) can be seen, in some sense, as an extension of that in Bradic *et al* (2011) from the parametric linear models to the flexible GSVCMs.

Our model selection and structure specification procedure is based on maximising the objective function in either (2.13) or (2.14). Let

$$\widehat{\boldsymbol{\alpha}}_j = \big(\widehat{\alpha}_{j1},\ \cdots,\ \widehat{\alpha}_{jn}\big)^{\mathrm{T}} \text{ and } \widehat{\boldsymbol{\beta}}_j = \big(\widehat{\beta}_{j1},\ \cdots,\ \widehat{\beta}_{jn}\big)^{\mathrm{T}},\ \ j = 1,\ \cdots,\ d_n, \tag{2.15}$$

be the maximiser of $\mathcal{Q}_n^1(\mathcal{A},\ \mathcal{B})$, and

$$\overline{\boldsymbol{\alpha}}_j = \big(\overline{\alpha}_{j1},\ \cdots,\ \overline{\alpha}_{jn}\big)^{\mathrm{T}} \text{ and } \overline{\boldsymbol{\beta}}_j = \big(\overline{\beta}_{j1},\ \cdots,\ \overline{\beta}_{jn}\big)^{\mathrm{T}},\ \ j = 1,\ \cdots,\ d_n, \tag{2.16}$$

be the maximiser of $\mathcal{Q}_n^2(\mathcal{A},\ \mathcal{B})$. The asymptotic theorems and remarks in Section 3 show that the estimators defined in (2.15) and (2.16) equal to the biased oracle estimators of the functional coefficients (see Section 3 for the definition) with probability approaching one.

2.4. *Estimation of the constant coefficients.* We next discuss how to estimate the constant coefficients in GSVCMs. By choosing the penalty function as the adaptive group LASSO (or the group SCAD) penalty, we would expect $\|\widehat{\boldsymbol{\alpha}}_j\| = 0$ (or $\|\overline{\boldsymbol{\alpha}}_j\| = 0$) when $a_j(\cdot) = 0$, and $\|\widehat{\boldsymbol{\beta}}_j\| = 0$ (or $\|\overline{\boldsymbol{\beta}}_j\| = 0$) when $a_j(\cdot)$ is a constant. Hence our model selection and structure specification procedure works as follows: if $\|\widehat{\boldsymbol{\alpha}}_j\| = 0$ (or $\|\overline{\boldsymbol{\alpha}}_j\| = 0$), the corresponding variable $x_j$ is not significant and should be removed from the model; if $\|\widehat{\boldsymbol{\beta}}_j\| = 0$ (or $\|\overline{\boldsymbol{\beta}}_j\| = 0$), the functional coefficient of $a_j(\cdot)$ is a constant which is denoted by $c_j$ and can be estimated by

$$\widehat{c}_j = n^{-1}\sum_{i=1}^{n} \widehat{\alpha}_{ji} \ \text{ or } \ \overline{c}_j = n^{-1}\sum_{i=1}^{n} \overline{\alpha}_{ji},\ \ j = s_{n1}+1, \cdots, s_{n2}. \tag{2.17}$$

Then the semi-varying coefficient modelling structure is finally specified.

**3. Asymptotic theory.** In this section, we present the asymptotic properties of the model selection and structure specification procedure introduced in Section 2. Recall that

$$\mathbf{a}_{k0} = \big[a_1(U_k),\ \cdots,\ a_{d_n}(U_k)\big]^{\mathrm{T}} \quad \text{and} \quad \mathbf{b}_{k0} = \big[\dot{a}_1(U_k),\ \cdots,\ \dot{a}_{d_n}(U_k)\big]^{\mathrm{T}}$$

for $k = 1, \cdots, n$. We start with the uniform consistency results for their penalised local maximum likelihood estimators

$$\widetilde{\mathbf{a}}_k = \big[\widetilde{a}_1(U_k),\ \cdots,\ \widetilde{a}_{d_n}(U_k)\big]^{\mathrm{T}} \quad \text{and} \quad \widetilde{\mathbf{b}}_k = \big[\widetilde{\dot{a}}_1(U_k), \cdots, \widetilde{\dot{a}}_{d_n}(U_k)\big]^{\mathrm{T}},$$

which are the maximisers of the objective function in (2.3). In the sequel, we let $\alpha_n \propto \beta_n$ denote $c_1\beta_n \le \alpha_n \le c_2\beta_n$ when $n$ is sufficiently large, where $0 < c_1 \le c_2 < \infty$.

**Proposition 3.1**. *Suppose that Assumptions 1–4 in Appendix A are satisfied.*
*(i) If the moment condition (A.1) and Assumption 5 are satisfied with $d_n \propto n^{\tau_1}$, $0 \le \tau_1 < \infty$, we have*

$$\max_{1\le k\le n} \|\widetilde{\mathbf{a}}_k - \mathbf{a}_{k0}\| + \max_{1\le k\le n} \|h(\widetilde{\mathbf{b}}_k - \mathbf{b}_{k0})\| = O_P(\sqrt{s_{n2}}\lambda_1), \tag{3.1}$$

*where $s_{n2}$ is the number of the non-zero functional coefficients*
*(ii) If the moment condition (A.2) and Assumption 5′ are satisfied with $d_n \propto \exp\{(nh)^{\tau_2}\}$, then (3.1) also holds, where $0 \le \tau_2 < 1 - \tau_3$ with $0 < \tau_3 < 1$.*

**Remark 3.1**. The above proposition indicates that the preliminary estimators $\widetilde{\mathbf{a}}_k$ and $\widetilde{\mathbf{b}}_k$ are uniformly consistent, as Assumption 3 in Appendix A guarantees that the maximal distance between two consecutive index variables $U_i$ is only with the order $O_P(\log n/n)$ (c.f., Janson 1987) and the observed values of $U$ can be sufficiently dense on the compact support $[0, 1]$. The uniform convergence rate in (3.1) depends on $s_{n2}$, the number of the non-zero functional coefficients, and the tuning parameter $\lambda_1$. In Assumptions 5 and 5′, we impose some condition on the relationship between $\lambda_1$ and the well-known uniform convergence rate $\big(\frac{\log h^{-1}}{nh}\big)^{1/2}$, and assume that $\lambda_1 \propto \lambda_2$. As a consequence, the influence of $\big(\frac{\log h^{-1}}{nh}\big)^{1/2}$ and $\lambda_2$ would be dominated by that of $\lambda_1$. Although the dimension of potential covariates in our model can be larger than the sample size and diverge at an exponential rate, $s_{n2}$ is not allowed to diverge too fast in order to guarantee the consistency of the preliminary estimators of the functional coefficients. The condition $s_{n2}\lambda_1^2 h^{-2} = o(1)$ in Assumptions 5 and 5′ indicates that $s_{n2}$ is

allowed to be divergent at a slow polynomial rate of $n$. It is also interesting to find from the comparison between (A.1) and (A.2) in Appendix A that the required moment condition when $d_n$ diverges at a polynomial rate is weaker than that when $d_n$ diverges at an exponential rate.

**Remark 3.2**. Note that in the penalised local log-likelihood estimation method in Section 2.1, we do not use the group LASSO or SCAD penalty function. Although Proposition 3.1 establishes the uniform consistency for the preliminary estimators of the functional coefficients and their derivatives, the shrinkage estimation method in Section 2.1 does not have the well-known sure screening property (Fan *et al*, 2014; Liu *et al*, 2014). However, under some further conditions, the uniform convergence rate in (3.1) would be sufficient for us to prove Theorems 3.1 and 3.2 below.

Let

$$\overline{\mathcal{A}}_n = \big(\overline{\mathbf{a}}_1^{\mathrm{T}}, \cdots, \overline{\mathbf{a}}_n^{\mathrm{T}}\big)^{\mathrm{T}} \quad \text{and} \quad \overline{\mathcal{B}}_n = \big(\overline{\mathbf{b}}_1^{\mathrm{T}}, \cdots, \overline{\mathbf{b}}_n^{\mathrm{T}}\big)^{\mathrm{T}},$$

where $\overline{\mathbf{a}}_k = (\overline{\alpha}_{1k}, \cdots, \overline{\alpha}_{d_nk})^{\mathrm{T}}$ and $\overline{\mathbf{b}}_k = (\overline{\beta}_{1k}, \cdots, \overline{\beta}_{d_nk})^{\mathrm{T}}$. Let $\mathbf{a}_k^o$ be any $d_n$-dimensional vector with the last $(d_n - s_{n2})$ elements being zeros, and $\mathbf{b}_k^o$ be any $d_n$-dimensional vector with the last $(d_n - s_{n1})$ elements being zeros. Denote

$$\mathcal{A}^o = \big[(\mathbf{a}_1^o)^{\mathrm{T}}, \cdots, (\mathbf{a}_n^o)^{\mathrm{T}}\big]^{\mathrm{T}} \quad \text{and} \quad \mathcal{B}^o = \big[(\mathbf{b}_1^o)^{\mathrm{T}}, \cdots, (\mathbf{b}_n^o)^{\mathrm{T}}\big]^{\mathrm{T}},$$

and then define the biased oracle estimators

$$\overline{\mathcal{A}}_n^{bo} = \big[(\overline{\mathbf{a}}_1^{bo})^{\mathrm{T}}, \cdots, (\overline{\mathbf{a}}_n^{bo})^{\mathrm{T}}\big]^{\mathrm{T}} \quad \text{and} \quad \overline{\mathcal{B}}_n^{bo} = \big[(\overline{\mathbf{b}}_1^{bo})^{\mathrm{T}}, \cdots, (\overline{\mathbf{b}}_n^{bo})^{\mathrm{T}}\big]^{\mathrm{T}},$$

which maximise the objective function $\mathcal{Q}_n^2(\mathcal{A}^o, \mathcal{B}^o)$ when the penalty function is the SCAD penalty. The following theorem gives the relation between the penalised estimators which maximise the objective function (2.14) and the corresponding biased oracle estimators when the SCAD penalty is used. The result for the case of the adaptive group LASSO penalty is similar and will be discussed in Remark 3.3 below.

**Theorem 3.1**. *Suppose that the conditions in Proposition 3.1(i) are satisfied. When the penalty functions are defined in (2.11) and (2.12), and Assumption 6 in Appendix A is satisfied, with probability approaching one, the maximiser of the objective function $\mathcal{Q}_n^2(\cdot,\cdot)$ defined in (2.14), $(\overline{\mathcal{A}}_n, \overline{\mathcal{B}}_n)$, exists and equals to $(\overline{\mathcal{A}}_n^{bo}, \overline{\mathcal{B}}_n^{bo})$. Furthermore,*

$$\frac{1}{n}\big\|\overline{\mathcal{A}}_n^{bo} - \mathcal{A}_0\big\|^2 = \frac{s_{n2}}{nh}, \quad \frac{1}{n}\big\|\overline{\mathcal{B}}_n^{bo} - \mathcal{B}_0\big\|^2 = \frac{s_{n2}}{nh^3}, \tag{3.2}$$

*where*

$$\mathcal{A}_0 = (\mathbf{a}_{10}^T, \cdots, \mathbf{a}_{n0}^T)^T, \quad \mathcal{B}_0 = (\mathbf{b}_{10}^T, \cdots, \mathbf{b}_{n0}^T)^T.$$

**Remark 3.3**. Given the moment condition (A.2) and Assumption 5′ in Appendix A with $d_n \propto \exp\{(nh)^{\tau_2}\}$, the above result still holds. It can be proved by using Proposition 3.1(ii) and strengthening (A.3) in Assumption 6 to

$$h^{-1/2}\left[\left(\frac{\log h^{-1}}{nh}\right)^{\tau_3/2}\sqrt{nh} + s_{n2}^{1/2}(1+\lambda_1\sqrt{nh})\right] = o(\lambda_4),$$

where $\tau_3$ is defined in Proposition 3.1(ii). Noting that the left hand side is controlled by $\lambda_1\sqrt{ns_{n2}}$, the above condition can be simplified to $\lambda_1\sqrt{ns_{n2}} = o(\lambda_4)$. Theorem 3.1 suggests, using the proposed model selection procedure, the zero coefficients can be estimated exactly as zeros, and the derivatives of the constant coefficients can also be estimated exactly as zeros, which indicates that the *sparsity* property holds for the proposed model selection procedure. Hence, our theorem complements some existing ultra-high dimensional sparsity results such as those derived by Bradic *et al* (2011), Fan and Lv (2011) and Lian (2012). Furthermore, for the penalty functions defined in (2.11) and (2.12), by Proposition 3.1 and Assumption 6, we may show that properties (i)–(iv) for the folded concave penalty function introduced by Fan *et al* (2014) are satisfied with probability approaching one. Hence, Theorem 3.1 can also be seen, in some sense, as a generalisation of Theorem 1 in Fan *et al* (2014). When the adaptive group LASSO penalty is used, by modifying the conditions in Assumption 6(i), we may show that the above sparsity result still holds and (3.2) is satisfied by replacing $\overline{\mathcal{A}}_n^{bo}$ and $\overline{\mathcal{B}}_n^{bo}$ by $\widehat{\mathcal{A}}_n^{bo}$ and $\widehat{\mathcal{B}}_n^{bo}$, the biased oracle estimators with the adaptive group LASSO penalty.

We next study the oracle property for the penalised estimators of the non-zero functional coefficients and constant coefficients. Let $a_j^{uo}(U_k)$, $j = 1, \cdots, s_{n1}$, $k = 1, \cdots, n$, be the (unbiased) oracle estimator of $a_j(U_k)$, and $c_j^{uo}$, $j = s_{n1}+1, \cdots, s_{n2}$, be the (unbiased) oracle estimator of the constant coefficient $c_j$. The (unbiased) oracle estimators are obtained by the standard estimation procedure for the GSVCMs, i.e., the maximisation of the objective function $\mathcal{L}_n^{\diamond}(\mathcal{A}^o, \mathcal{B}^o)$ with respect to $\mathcal{A}^o$ and $\mathcal{B}^o$ (the penalty terms in (2.13) and (2.14) are ignored) and the application of (2.17) under the assumption that we know $a_j(\cdot) \equiv 0$ when $j = s_{n2}+1, \cdots, d_n$ and $a_j(\cdot) \equiv c_j$ when $j = s_{n1}+1, \cdots, s_{n2}$. In the following theorem, we only consider the case that the penalty functions are defined in (2.11) and (2.12) to save the

space. Let

$$\overline{\mathbf{D}}_n = \Big( \max_{1\le k\le n} \big|\overline{a}_1(U_k) - a_1^{uo}(U_k)\big|, \ \cdots, \ \max_{1\le k\le n} \big|\overline{a}_{s_{n1}}(U_k) - a_{s_{n1}}^{uo}(U_k)\big| \Big)^{\mathrm{T}},$$

where $\overline{a}_j(U_k) = \overline{\alpha}_{jk}$ is defined in (2.16), and

$$\mathbf{C}_n^{uo} = \big(c_{s_{n1}+1}^{uo}, \cdots, c_{s_{n2}}^{uo}\big)^{\mathrm{T}}, \quad \overline{\mathbf{C}}_n = \big(\overline{c}_{s_{n1}+1}, \cdots, \overline{c}_{s_{n2}}\big)^{\mathrm{T}},$$

where $\overline{c}_j$ is defined in (2.17).

**Theorem 3.2**. *Suppose that the conditions of Theorem 3.1 are satisfied. For any $s_{n1}$-dimensional vector $\mathbf{B}_n$ with $\|\mathbf{B}_n\| = 1$, we have*

$$\sqrt{nh}\mathbf{B}_n^T\overline{\mathbf{D}}_n = o_P(1); \tag{3.3}$$

*and for any $(s_{n2} - s_{n1})$-dimensional vector $\mathbf{A}_n$ with $\|\mathbf{A}_n\| = 1$, we have*

$$\sqrt{n}\mathbf{A}_n^T\left(\overline{\mathbf{C}}_n - \mathbf{C}_n^{uo}\right) = o_P(1). \tag{3.4}$$

**Remark 3.4**. Theorem 3.2 indicates that the penalised likelihood estimators of the non-zero functional coefficients and constant coefficients have the same asymptotic distribution as the corresponding oracle estimators, and thus the oracle property holds. As discussed in Remark 3.3, by strengthening the moment conditions, we can also show that the above oracle property holds when $d_n \propto \exp\big\{(nh)^{\tau_2}\big\}$. Following the arguments in Zhang and Peng (2010) and Li *et al* (2013), we can easily establish the asymptotic normality of $\overline{a}_j(\cdot)$, $j = 1, \cdots, s_{n1}$, and $\overline{c}_j$, $j = s_{n1}+1, \cdots, s_{n2}$.

**4. Computational algorithm and selection of tuning parameters.** In this section, we introduce a computational algorithm to maximise $\mathcal{Q}_n^1(\mathcal{A},\ \mathcal{B})$ and $\mathcal{Q}_n^2(\mathcal{A},\ \mathcal{B})$ defined in Section 2.3 and discuss how to choose the tuning parameters involved in the proposed penalised likelihood method.

4.1. *Computational algorithm.* We first re-arrange the quadratic objective function $\mathcal{L}_n^{\diamond}(\mathcal{A},\ \mathcal{B})$ in order to make it have the standard form when using the penalised estimation method. Let

$$\boldsymbol{\theta} = \big(\boldsymbol{\alpha}_1^{\mathrm{T}}, \ \cdots, \ \boldsymbol{\alpha}_{d_n}^{\mathrm{T}}, \ h\boldsymbol{\beta}_1^{\mathrm{T}}, \ \cdots, \ h\boldsymbol{\beta}_{d_n}^{\mathrm{T}}\big)^{\mathrm{T}}$$

and define the transformation matrix

$$\mathbf{T} = (I_n \otimes e_{1,2d_n}, \ \cdots, \ I_n \otimes e_{d_n,2d_n}, \ I_n \otimes e_{d_n+1,2d_n} \ \cdots, \ I_n \otimes e_{2d_n,2d_n})^{\mathrm{T}},$$

where $e_{k,d}$ is a $d$-dimensional unit vector with the $k$th component being 1 and $I_n$ is an $n \times n$ identity matrix. With the above notations, it is easy to show that $\boldsymbol{\theta} = \mathbf{T}\mathcal{V}_n(\mathcal{A}, h\mathcal{B})$, where $\mathcal{V}_n(\mathcal{A}, h\mathcal{B})$ is defined as in Section 2.2. Let $\tilde{\boldsymbol{\theta}}$ be defined as $\boldsymbol{\theta}$ but with $\mathcal{A}$ and $\mathcal{B}$ replaced by $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}$, respectively, and

$$\mathbf{H}^2 = \mathbf{H}^{\mathrm{T}}\mathbf{H} = -\mathbf{T}\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n)\mathbf{T}^{\mathrm{T}}, \quad \tilde{\boldsymbol{\eta}} = \mathbf{H}\tilde{\boldsymbol{\theta}} + (\mathbf{H}^{-1})^{\mathrm{T}}\mathbf{T}\dot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n).$$

We define a quadratic objective function

$$\mathcal{L}_n^*(\mathcal{A},\ \mathcal{B}) = -\frac{1}{2}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta})^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}).$$

Given the initial estimator $\mathcal{V}_n(\widetilde{\mathcal{A}}_n, h\widetilde{\mathcal{B}}_n)$, it is easy to see the difference between $\mathcal{L}_n^\diamond(\mathcal{A},\ \mathcal{B})$ and $\mathcal{L}_n^*(\mathcal{A},\ \mathcal{B})$ is a constant. Therefore, the maximiser of $\mathcal{Q}_n^1(\mathcal{A},\ \mathcal{B})$ or $\mathcal{Q}_n^2(\mathcal{A},\ \mathcal{B})$ is the minimiser of the following target function:

$$\mathcal{O}(\boldsymbol{\theta}) \equiv \frac{1}{2}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta})^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}) + \sum_{j=1}^{d_n} \tau_{1j}\|\boldsymbol{\alpha}_j\| + \sum_{j=1}^{d_n} \tau_{2j}\|h\boldsymbol{\beta}_j\|, \tag{4.1}$$

where

$$\tau_{1j} = \lambda_3\|\widetilde{\boldsymbol{\alpha}}_j\|^{-\kappa} \text{ and } \tau_{2j} = \lambda_3^*\widetilde{D}_j^{-\kappa}$$

for $\mathcal{Q}_n^1(\mathcal{A},\ \mathcal{B})$; and

$$\tau_{1j} = \dot{p}_{\lambda_4}(\|\widetilde{\boldsymbol{\alpha}}_j\|) \text{ and } \tau_{2j} = \dot{p}_{\lambda_4^*}(\widetilde{D}_j)$$

for $\mathcal{Q}_n^2(\mathcal{A},\ \mathcal{B})$. As a direct consequence of the Karush-Kuhn-Tucker conditions, we have that a necessary and sufficient condition for $\boldsymbol{\theta}$ to be a minimiser of $\mathcal{O}(\boldsymbol{\theta})$ is, for $j = 1, \cdots, d_n$,

$$\begin{cases} -\mathrm{H}_j^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}) + \tau_{1j}\|\boldsymbol{\alpha}_j\|^{-1}\boldsymbol{\alpha}_j = \mathbf{0}_n & \forall\ \boldsymbol{\alpha}_j \neq \mathbf{0}_n, \\ \|\mathrm{H}_j^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta})\| < \tau_{1j} & \forall\ \boldsymbol{\alpha}_j = \mathbf{0}_n, \\ -\mathrm{H}_{j+d_n}^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}) + \tau_{2j}\|\boldsymbol{\beta}_j\|^{-1}\boldsymbol{\beta}_j = \mathbf{0}_n & \forall\ \boldsymbol{\beta}_j \neq \mathbf{0}_n, \\ \|\mathrm{H}_{j+d_n}^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta})\| < \tau_{2j} & \forall\ \boldsymbol{\beta}_j = \mathbf{0}_n, \end{cases}$$

where $\mathrm{H}_j$ is the matrix consisting of the $((j-1)n+1)$-th to the $(jn)$-th column of $\mathbf{H}$ and $\mathbf{0}_n$ is an $n$-dimensional vector with each component being 0. Hence, for $j = 1, \cdots, d_n$, we have $\boldsymbol{\alpha}_j = \mathbf{0}_n$ if $\|\mathrm{H}_j^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}_{-j})\| < \tau_{1j}$, otherwise

$$\boldsymbol{\alpha}_j = \left(\mathrm{H}_j^{\mathrm{T}}\mathrm{H}_j + \tau_{1j}\|\boldsymbol{\alpha}_j\|^{-1}I_n\right)^{-1}\mathrm{H}_j^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}_{-j});$$

and $\boldsymbol{\beta}_j = \mathbf{0}_n$ if $\|\mathrm{H}_{j+d_n}^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}_{-(j+d_n)})\| < \tau_{2j}$, otherwise

$$\boldsymbol{\beta}_j = \left(h\mathrm{H}_{j+d_n}^{\mathrm{T}}\mathrm{H}_{j+d_n} + \tau_{2j}\|\boldsymbol{\beta}_j\|^{-1}I_n\right)^{-1}\mathrm{H}_{j+d_n}^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}_{-(j+d_n)}),$$

where

$$\begin{aligned}\boldsymbol{\theta}_{-j} =&(\boldsymbol{\alpha}_1^{\mathrm{T}},\ \cdots,\ \boldsymbol{\alpha}_{j-1}^{\mathrm{T}},\ \mathbf{0}_n^{\mathrm{T}},\ \boldsymbol{\alpha}_{j+1}^{\mathrm{T}},\ \cdots,\ \boldsymbol{\alpha}_{d_n}^{\mathrm{T}},\ h\boldsymbol{\beta}_1^{\mathrm{T}},\ \cdots,\ h\boldsymbol{\beta}_{d_n}^{\mathrm{T}})^{\mathrm{T}},\\ \boldsymbol{\theta}_{-(j+d_n)} =&(\boldsymbol{\alpha}_1^{\mathrm{T}},\ \cdots,\ \boldsymbol{\alpha}_{d_n}^{\mathrm{T}},\ h\boldsymbol{\beta}_1^{\mathrm{T}},\ \cdots,\ h\boldsymbol{\beta}_{j-1}^{\mathrm{T}},\ \mathbf{0}_n^{\mathrm{T}},\ h\boldsymbol{\beta}_{j+1}^{\mathrm{T}},\ \cdots,\ h\boldsymbol{\beta}_{d_n}^{\mathrm{T}})^{\mathrm{T}}.\end{aligned}$$

This leads to the following iterative algorithm to obtain the minimisers of $\mathcal{O}(\boldsymbol{\theta})$.

**Step 1.** Start with $\boldsymbol{\alpha}_j^{(0)} = \widetilde{\boldsymbol{\alpha}}_j$ and $\boldsymbol{\beta}_j^{(0)} = \widetilde{\boldsymbol{\beta}}_j$, $j = 1,\ \cdots,\ d_n$, where $\widetilde{\boldsymbol{\alpha}}_j$ and $\widetilde{\boldsymbol{\beta}}_j$ are the preliminary estimates of the functional coefficients $[a_j(U_1),\cdots,a_j(U_n)]^{\mathrm{T}}$ and their derivatives $[\dot{a}_j(U_1),\cdots,\dot{a}_j(U_n)]^{\mathrm{T}}$, respectively, which are introduced in Section 2.1.

**Step 2.** For $j = 1,\ \cdots,\ d_n$, let $\boldsymbol{\alpha}_j^{(k)}$ and $\boldsymbol{\beta}_j^{(k)}$ be the results after the $k$-th iteration. Update $\boldsymbol{\alpha}_j^{(k)}$ and $\boldsymbol{\beta}_j^{(k)}$ in the $(k+1)$-th iteration as follows: for $j = 1,\ \cdots,\ d_n$, $\boldsymbol{\alpha}_j^{(k+1)} = \mathbf{0}_n$ if $\|\mathrm{H}_j^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}_{-j}^{(k)})\| < \tau_{1j}^{(k)}$, otherwise

$$\boldsymbol{\alpha}_j^{(k+1)} = \Big(\mathrm{H}_j^{\mathrm{T}}\mathrm{H}_j + \tau_{1j}^{(k)}\|\boldsymbol{\alpha}_j^{(k)}\|^{-1}I_n\Big)^{-1}\mathrm{H}_j^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}_{-j}^{(k)});$$

and $\boldsymbol{\beta}_j^{(k+1)} = \mathbf{0}_n$ if $\|\mathrm{H}_{j+d_n}^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}_{-(j+d_n)}^{(k)})\| < \tau_{2j}^{(k)}$, otherwise

$$\boldsymbol{\beta}_j^{(k+1)} = \Big(h\mathrm{H}_{j+d_n}^{\mathrm{T}}\mathrm{H}_{j+d_n} + \tau_{2j}^{(k)}\|\boldsymbol{\beta}_j^{(k)}\|^{-1}I_n\Big)^{-1}\mathrm{H}_{j+d_n}^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}_{-(j+d_n)}^{(k)});$$

where $\tau_{1j}^{(k)}$ is defined as $\tau_{1j}$ in (4.1) but with $\widetilde{\boldsymbol{\alpha}}_j$ replaced by $\boldsymbol{\alpha}_j^{(k)}$, $\tau_{2j}^{(k)}$ is defined as $\tau_{2j}$ in (4.1) but with $\widetilde{D}_j$ replaced by $D_j^{(k)}$,

$$\begin{aligned}D_j^{(k)} =&\Big\{\sum_{s=1}^{n}\big[a_j^{(k)}(U_s) - \frac{1}{n}\sum_{l=1}^{n}a_j^{(k)}(U_l)\big]^2\Big\}^{1/2},\\ \boldsymbol{\theta}_{-j}^{(k)} =&\Big[(\boldsymbol{\alpha}_1^{(k+1)})^{\mathrm{T}},\cdots,(\boldsymbol{\alpha}_{j-1}^{(k+1)})^{\mathrm{T}},\mathbf{0}_n^{\mathrm{T}},(\boldsymbol{\alpha}_{j+1}^{(k)})^{\mathrm{T}},\cdots,(\boldsymbol{\alpha}_{d_n}^{(k)})^{\mathrm{T}},\\ &(h\boldsymbol{\beta}_1^{(k)})^{\mathrm{T}},\cdots,(h\boldsymbol{\beta}_{d_n}^{(k)})^{\mathrm{T}}\Big]^{\mathrm{T}},\quad\text{and}\\ \boldsymbol{\theta}_{-(j+d_n)}^{(k)} =&\Big[(\boldsymbol{\alpha}_1^{(k+1)})^{\mathrm{T}},\cdots,(\boldsymbol{\alpha}_{d_n}^{(k+1)})^{\mathrm{T}},(h\boldsymbol{\beta}_1^{(k+1)})^{\mathrm{T}},\cdots,(h\boldsymbol{\beta}_{j-1}^{(k+1)})^{\mathrm{T}},\mathbf{0}_n^{\mathrm{T}},\\ &(h\boldsymbol{\beta}_{j+1}^{(k)})^{\mathrm{T}},\cdots,(h\boldsymbol{\beta}_{d_n}^{(k)})^{\mathrm{T}}\Big]^{\mathrm{T}}.\end{aligned}$$

Furthermore, if $\|\boldsymbol{\alpha}_j^{(k)}\| = \mathbf{0}_n$ and $\|\mathrm{H}_j^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}_{-j}^{(k)})\| > \tau_{1j}^{(k)}$, we set

$$\boldsymbol{\alpha}_j^{(k+1)} = \Big(\mathrm{H}_j^{\mathrm{T}}\mathrm{H}_j + (\tau_{1j}^{(k)}/\Delta_{\boldsymbol{\alpha}}^{(k)})I_n\Big)^{-1}\mathrm{H}_j^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}_{-j}^{(k)})$$

with $\Delta_{\boldsymbol{\alpha}}^{(k)} = \min\Big\{\|\boldsymbol{\alpha}_l^{(k)}\| : \ \|\boldsymbol{\alpha}_l^{(k)}\| \neq 0, \ l = 1, \cdots, d_n\Big\}$. If $\|\boldsymbol{\beta}_j^{(k)}\| = \mathbf{0}_n$ and $\|\mathrm{H}_{j+d_n}^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{-(j+d_n)}^{(k)})\| > \tau_{2j}^{(k)}$, we set

$$\boldsymbol{\beta}_j^{(k+1)} = \Big(h\mathrm{H}_{j+d_n}^{\mathrm{T}}\mathrm{H}_{j+d_n} + (\tau_{2j}^{(k)}/\Delta_{\boldsymbol{\beta}}^{(k)})I_n\Big)^{-1}\mathrm{H}_{j+d_n}^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}_{-(j+d_n)}^{(k)})$$

with $\Delta_{\boldsymbol{\beta}}^{(k)} = \min\Big\{\|\boldsymbol{\beta}_l^{(k)}\| : \ \|\boldsymbol{\beta}_l^{(k)}\| \neq 0, \ l = 1, \ \cdots, \ d_n\Big\}$.

**Step 3.** If $\sum_{j=1}^{d_n}\Big(\|\boldsymbol{\alpha}_j^{(k)} - \boldsymbol{\alpha}_j^{(k+1)}\| + h\|\boldsymbol{\beta}_j^{(k)} - \boldsymbol{\beta}_j^{(k+1)}\|\Big)$ is smaller than a chosen threshold, we stop the iteration, and $\big(\boldsymbol{\alpha}_j^{(k+1)}, \ \boldsymbol{\beta}_j^{(k+1)}\big), j = 1, \cdots, d_n$, are the minimisers of $\mathcal{O}(\boldsymbol{\theta})$.

The simulation studies in Section 5 below will show that the above iterative procedure works reasonably well in the finite sample cases. The simulation studies are conducted by a small computer cluster which contains 64 CPUs while the real data analysis results are obtained by a single PC within one day.

4.2. *Selection of the tuning parameters.* The tuning parameters involved in the proposed model selection and structure specification procedure play a very important role. We next discuss how to choose these tuning parameters. First, for the preliminary estimates, the tuning parameters $\lambda_1$ and $\lambda_2$ are selected through BIC, and the bandwidth is set to be $h = 0.75[(\log d_n)/n]^{0.2}$. The reason for not using a data-driven method to select the bandwidth $h$ is to reduce the computational cost. Also the preliminary estimation is not very sensitive to the choice of the bandwidth. Then, for the model selection and specification procedure based on $\mathcal{Q}_n^1(\mathcal{A}, \ \mathcal{B})$ or $\mathcal{Q}_n^2(\mathcal{A}, \ \mathcal{B})$, the tuning parameters $\lambda_3$ and $\lambda_3^*$ or $\lambda_4$ and $\lambda_4^*$ are selected by the generalized information criterion (GIC) proposed by Fan and Tang (2013). We next briefly introduce the GIC method.

As the models concerned involve both unknown constant parameters and unknown functional parameters, to use GIC, we first need to figure out how many unknown constant parameters an unknown functional parameter amounts to. Cheng *et al* (2009) suggest that an unknown functional parameter would amount to $1.028571h^{-1}$ unknown constant parameters when Epanechnikov kernel is used. Hence we construct the GIC for model (2.1) as

$$\begin{aligned}\mathrm{GIC}(\lambda, \lambda^*) \ &= \ -2\sum_{i=1}^{n}\ell(\hat{m}(U_i, X_i), \ y_i) + \\ &\quad 2\ln\{\ln(n)\}\ln(1.028571d_nh^{-1})(k_1 + 1.028571k_2h^{-1}),\end{aligned}$$

where $\hat{m}(U_i, X_i)$ is defined as $m(U_i, X_i)$ with all unknowns being replaced by their estimators obtained based on the tuning parameters $\lambda_3$ and $\lambda_3^*$ (or $\lambda_4$ and $\lambda_4^*$), $k_1$ is the number of significant covariates with constant coefficients obtained based on the given pair of tuning parameters, and $k_2$ is the number of significant covariates with functional coefficients obtained based on the given pair of tuning parameters. For the maximisation of $\mathcal{Q}_n^1(\mathcal{A},\ \mathcal{B})$, the minimiser of $\mathrm{GIC}(\lambda_3, \lambda_3^*)$ is the selected $\lambda_3$ and $\lambda_3^*$, while for the maximisation of $\mathcal{Q}_n^2(\mathcal{A},\ \mathcal{B})$, the minimiser of $\mathrm{GIC}(\lambda_4, \lambda_4^*)$ is the selected $\lambda_4$ and $\lambda_4^*$.

**5. Simulation studies.** In this section, we give three simulated examples to examine the accuracy of the proposed model selection, structure specification and estimation procedure, as well as the oracle property of the proposed estimators. Throughout this section, we call the procedure based on (2.13) the adaptive group LASSO method and the procedure based on (2.14) the group SCAD method. For the adaptive group LASSO method, the pre-determined parameter $\kappa$ is chosen to be 1. For the group SCAD method, the SCAD penalty is defined through its derivative as in (2.10). The kernel function used in this section is taken to be the Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$. The bandwidth and other tuning parameters are selected by the approach described in Section 4.2.

We will start with a simulated example on a semi-varying coefficient Poisson regression model, then an example on varying coefficient models and finally an example on a varying coefficient logistic regression model. In Example 5.1, we will compare the performance of the proposed adaptive group LASSO and group SCAD methods on model selection, structure specification and estimation, and find that the group SCAD method gives slightly better finite sample performance under all simulation settings. Thus we will call the group SCAD method "our method" in the following two examples and only compare it with some existing methods. In Example 5.2, we will compare our method with the KLASSO proposed in Wang and Xia (2009) based on varying coefficient models. In Example 5.3, we will compare our method with the method proposed in Lian (2012). The simulation results of the KLASSO and Lian's method in Tables 3-5 are the original results reported in Wang and Xia (2009) and Lian (2012), respectively. From the simulation results, we will find that our method outperforms the existing ones.

**Example 5.1.** We generate a sample from a Poisson regression model as follows: first independently generate $x_{ij}$, $i = 1, \cdots, n$, $j = 1, \cdots, d_n$, from

the standard normal distribution $\mathsf{N}(0,\ 1)$, and $U_i$, $i = 1, \cdots, n$, from uniform distribution $\mathsf{U}[0,\ 1]$, and then generate $y_i$ based on

$$\mathsf{P}(y_i = k) = \frac{\xi_i^k}{k!} e^{-\xi_i}, \quad \log(\xi_i) = \sum_{j=1}^{d_n} a_j(U_i) x_{ij}. \tag{5.1}$$

We set the sample size $n = 200$, the number of significant covariates $s_{n2}$ to be the integer part of $\ln n$ and $a_j(\cdot)$s in (5.1) to be

$$a_1(U) = -U, \quad a_2(U) = \sin(2\pi U), \quad a_3(U) = 4(U - 0.5)^2$$
$$a_4(U) = c_1 = 0.6, \quad a_5(U) = c_2 = -0.7, \quad a_j(U) = 0 \text{ for } j > 5.$$

For dimensions $d_n = 50$, $d_n = 100$, $d_n = 200$, and $d_n = 500$, we apply both the adaptive group LASSO method and the group SCAD method to the simulated sample to select the model, and estimate the unknown functional or constant coefficients. For each case, we do $1,000$ simulations, and compute the mean integrated squared error (MISE) of the estimators of the unknown functional coefficients and the mean squared error (MSE) of the estimators of the unknown constant coefficients. We also calculate the ratios of correct, under-selected, under-specified, over-selected, over-specified and other models. The "under-selected models" means the selected models ignoring the significant covariates. The "under-specified models" means where the functional coefficients are mis-specified as the constant coefficients. The "over-selected models" means the selected models including the insignificant covariates. The "over-specified models" means where the constant coefficients are mis-specified as functional. The "other models" means that there exist more than one incorrect situation as listed above. The "correct models" need not only select the true model but also correctly identify the modelling structure.

The simulation results are reported in Tables 1 and 2. We can see from Table 1 that both the adaptive group LASSO method and the group SCAD method work well for model selection and structure specification, and the group SCAD method gives slightly better performance. Table 2 shows that the estimators obtained by either the adaptive group LASSO method or the group SCAD method are doing very well, and their performance is comparable to that of the oracle estimators.

**Example 5.2.** As the varying coefficient models are a special case of the generalised varying coefficient models, our method is also applicable to the varying coefficient models. In this example, we compare our method with the

TABLE 1
*The ratios of model selection in 1,000 simulations*

| Adaptive group LASSO method | | | | | | |
|---|---|---|---|---|---|---|
| $d_n$ | Correct | Under-selected | Under-specified | Over-selected | Over-specified | Others |
| 50 | 0.967 | 0.001 | 0.002 | 0.005 | 0.024 | 0.001 |
| 100 | 0.944 | 0.001 | 0.005 | 0.008 | 0.039 | 0.003 |
| 200 | 0.915 | 0.006 | 0.014 | 0.012 | 0.045 | 0.008 |
| 500 | 0.863 | 0.015 | 0.023 | 0.026 | 0.057 | 0.016 |
| Group SCAD method | | | | | | |
| $d_n$ | Correct | Under-selected | Under-specified | Over-selected | Over-specified | Others |
| 50 | 0.970 | 0.001 | 0.002 | 0.003 | 0.022 | 0.002 |
| 100 | 0.948 | 0.002 | 0.004 | 0.006 | 0.038 | 0.002 |
| 200 | 0.925 | 0.005 | 0.012 | 0.010 | 0.042 | 0.006 |
| 500 | 0.878 | 0.013 | 0.020 | 0.022 | 0.051 | 0.016 |

*The ratios of choosing the correct, under-selected, under-specified, over-selected, over-specified and other models in 1000 simulations by using either the adaptive group LASSO method or the group SCAD method.*

TABLE 2
*The MISEs and MSEs of the estimators for the functional and constant coefficients*

| | Adaptive group LASSO | | | Group SCAD | | | Oracle Estimators | | |
|---|---|---|---|---|---|---|---|---|---|
| $d_n$ | $\widehat{a}_1(\cdot)$ | $\widehat{a}_2(\cdot)$ | $\widehat{a}_3(\cdot)$ | $\overline{a}_1(\cdot)$ | $\overline{a}_2(\cdot)$ | $\overline{a}_3(\cdot)$ | $a_1^{uo}(\cdot)$ | $a_2^{uo}(\cdot)$ | $a_3^{uo}(\cdot)$ |
| 50 | 0.026 | 0.037 | 0.038 | 0.024 | 0.033 | 0.036 | 0.019 | 0.023 | 0.025 |
| 100 | 0.038 | 0.049 | 0.050 | 0.035 | 0.045 | 0.048 | 0.019 | 0.023 | 0.025 |
| 200 | 0.058 | 0.069 | 0.072 | 0.052 | 0.063 | 0.066 | 0.019 | 0.023 | 0.025 |
| 500 | 0.090 | 0.095 | 0.098 | 0.084 | 0.093 | 0.091 | 0.019 | 0.023 | 0.025 |
| $d_n$ | $\widehat{c}_1$ | $\widehat{c}_2$ | | $\overline{c}_1$ | $\overline{c}_2$ | | $c_1^{uo}$ | $c_2^{uo}$ | |
| 50 | 0.015 | 0.017 | | 0.014 | 0.017 | | 0.006 | 0.008 | |
| 100 | 0.020 | 0.022 | | 0.018 | 0.021 | | 0.006 | 0.008 | |
| 200 | 0.030 | 0.035 | | 0.025 | 0.029 | | 0.006 | 0.008 | |
| 500 | 0.046 | 0.050 | | 0.039 | 0.042 | | 0.006 | 0.008 | |

*The MISEs or MSEs of the estimators obtained by either the adaptive group LASSO method or the group SCAD method. For $j = 1, 2, 3$ and $k = 1, 2$, $\widehat{a}_j(\cdot)$s and $\widehat{c}_k$s are the estimators obtained by the adaptive group LASSO method, $\overline{a}_j(\cdot)$s and $\overline{c}_k$s are the estimators obtained by the group SCAD method, and $a_j^{uo}(\cdot)$s and $c_k^{uo}$s are the unbiased oracle estimators.*

KLASSO method proposed in Wang and Xia (2009) for varying coefficient models. We consider exactly the same simulated example as that in Wang and Xia (2009), that is the following three varying coefficient models:

(I) $y_i = 2\sin(2\pi U_i)x_{i1} + 4U_i(1-U_i)x_{i2} + \sigma\epsilon_i$,
(II) $y_i = \exp(2U_i - 1)x_{i1} + 8U_i(1-U_i)x_{i2} + 2\cos^2(2\pi U_i)x_{i3} + \sigma\epsilon_i$,

(III) $y_i = 4U_i x_{i1} + 2\sin(2\pi U_i)x_{i2} + x_{i3} + \sigma\epsilon_i,$

where $x_{i1} = 1$ for any $i$, $(x_{i2}, \cdots, x_{i7})^{\mathrm{T}}$ and $\epsilon_i$, $i = 1, \cdots, n$, are independently generated from a multivariate normal distribution with $\mathrm{cov}(x_{ij_1}, x_{ij_2}) = 0.5^{|j_1 - j_2|}$ for any $2 \le j_1, j_2 \le 7$ and the standard normal distribution $\mathsf{N}(0, 1)$, respectively, $U_i$, $i = 1, \cdots, n$, are independently generated from either uniform distribution $\mathsf{U}[0, 1]$ or Beta distribution $\mathsf{B}(4, 1)$, $\sigma$ is set to be 1.5. For each model, we conduct 200 simulations, and in each simulation, we apply either our method or the KLASSO to do model selection and estimation and then make the comparison. We measure the performance of model selection by reporting the percentages of correct-, under- and over-fitting. The obtained results are presented in Table 3. From Table 3, we can see our method performs better than the KLASSO in model selection.

As in Wang and Xia (2009), we employ the median of the relative estimation errors (MREE), obtained in the 200 simulations, to assess the accuracy of an estimation method. The relative estimation error (REE) is defined as

$$\mathrm{REE} = 100 \times \frac{\sum_{i=1}^{n}\sum_{j=1}^{d_n} |\hat{a}_j(U_i) - a_j(U_i)|}{\sum_{i=1}^{n}\sum_{j=1}^{d_n} \left|\hat{a}_j^{uo}(U_i) - a_j(U_i)\right|} \tag{5.2}$$

where $\hat{a}_j(\cdot)$ is the estimator of $a_j(\cdot)$, obtained by the estimation method concerned, and $\hat{a}_j^{uo}(\cdot)$ is the oracle estimator of $a_j(\cdot)$. The median of REEs of our method and the KLASSO under different situations are presented in Table 4, which shows our method is more accurate than the KLASSO on estimation side. We thus conclude that our method performs better than the KLASSO on both model selection and estimation.

**Example 5.3.** In this example, we compare the model selection performance of our method with the method proposed in Lian (2012) for generalised varying coefficient models. We consider exactly the same simulation settings as that in Example 2 of Lian (2012), that is the following varying coefficient logistic regression model where the conditional mean function is:

$$\mathsf{E}[y_i|X_i] = \frac{\exp\Big\{\sum_{j=1}^{d_n} a_j(U_i)x_{ij}\Big\}}{1 + \exp\Big\{\sum_{j=1}^{d_n} a_j(U_i)x_{ij}\Big\}}. \tag{5.3}$$

The covariates are generated as following: for any $i = 1, \cdots, n$, $x_{i1} = 1$ and $(x_{i2}, \cdots, x_{id_n})^{\mathrm{T}}$ are generated from a multivariate normal distribution with

TABLE 3
*Comparison of model selection between our method and KLASSO*

| | | Our Method | | | KLASSO | | |
|---|---|---|---|---|---|---|---|
| $f_U(\cdot)$ | $n$ | Under | Correct | Over | Under | Correct | Over |
| | | | Model I | | | | |
| U[0,1] | 100 | 0.020 | 0.910 | 0.070 | 0.09 | 0.74 | 0.16 |
| | 200 | 0.005 | 0.985 | 0.010 | 0.02 | 0.95 | 0.03 |
| B[4, 1] | 100 | 0.020 | 0.875 | 0.105 | 0.21 | 0.58 | 0.21 |
| | 200 | 0.005 | 0.950 | 0.045 | 0.08 | 0.86 | 0.05 |
| | | | Model II | | | | |
| U[0, 1] | 100 | 0.015 | 0.915 | 0.070 | 0.01 | 0.83 | 0.16 |
| | 200 | 0.005 | 0.990 | 0.005 | 0.00 | 0.99 | 0.01 |
| B[4, 1] | 100 | 0.015 | 0.890 | 0.095 | 0.01 | 0.82 | 0.18 |
| | 200 | 0.005 | 0.970 | 0.025 | 0.00 | 0.96 | 0.04 |
| | | | Model III | | | | |
| U[0, 1] | 100 | 0.010 | 0.935 | 0.055 | 0.02 | 0.85 | 0.13 |
| | 200 | 0.000 | 0.995 | 0.005 | 0.00 | 0.99 | 0.01 |
| B[4, 1] | 100 | 0.015 | 0.895 | 0.090 | 0.02 | 0.79 | 0.19 |
| | 200 | 0.005 | 0.975 | 0.020 | 0.00 | 0.96 | 0.04 |

*The columns corresponding to "Under", "Correct" and "Over" are the ratios of under-fitting, correct-fitting and over-fitting for our method and KLASSO under different situations.*

TABLE 4
*Comparison of estimation between our method and KLASSO*

| *Median of Relative Estimation Errors* | | | | | | | |
|---|---|---|---|---|---|---|---|
| $f_U(\cdot)$ | $n$ | Our Method | KLASSO | $f_U(\cdot)$ | $n$ | Our Method | KLASSO |
| | | | Model I | | | | |
| U[0,1] | 100 | 109.35 | 121.00 | B[4, 1] | 100 | 114.41 | 127.42 |
| U[0,1] | 200 | 101.78 | 115.45 | B[4, 1] | 200 | 103.49 | 122.12 |
| | | | Model II | | | | |
| U[0, 1] | 100 | 107.81 | 109.45 | B[4, 1] | 100 | 115.17 | 111.06 |
| U[0, 1] | 200 | 101.51 | 109.46 | B[4, 1] | 200 | 103.73 | 108.07 |
| | | | Model III | | | | |
| U[0, 1] | 100 | 106.71 | 116.53 | B[4, 1] | 100 | 112.39 | 118.91 |
| U[0, 1] | 200 | 101.21 | 110.59 | B[4, 1] | 200 | 104.06 | 113.43 |

$\text{cov}(x_{ij_1},\ x_{ij_2}) = 0.1^{|j_1-j_2|}$ for any $2 \le j_1,\ j_2 \le d_n$. The index variable $U_i$, $i = 1,\ \cdots,\ n$, are independently generated from the uniform distribution U[0, 1]. We set the $a_j(\cdot)$s in (5.3) to be

$$a_1(U) = -4(U^3 + 2U^2 - 2U), \quad a_2(U) = 4\cos(2\pi U),$$
$$a_3(U) = 3\exp\{U - 0.5\}, \quad a_j(U) = 0, \text{ when } j > 3.$$

Similar to Example 2 of Lian (2012), we set the sample size $n = 150$ and

dimension $d_n = 50$ or $d_n = 200$. For each case, the simulation results are based on 100 replications. The model selection performance is measured by the average number of correct and incorrect varying coefficients. The former one means the average number of significant covariates that are correctly selected into the final model while the latter means the average number of insignificant covariates that are falsely selected as significant. The comparison results are shown in Table 5, from which we can see our method gives better model selection results.

TABLE 5
*Comparison between our method and Lian's methods on model selection*

| Method | Average # of varying coef. Correct | Incorrect |
|---|---|---|
| | $d_n = 50$ | |
| GL(BIC) | 3 | 18.75 |
| GL(eBIC) | 3 | 16.33 |
| AGL(BIC-BIC) | 3 | 10.29 |
| AGL(eBIC-eBIC) | 3 | 1.56 |
| Our Method | 3 | 1.37 |
| | $d_n = 200$ | |
| GL(BIC) | 3 | 38.78 |
| GL(eBIC) | 3 | 21.04 |
| AGL(BIC-BIC) | 3 | 25.72 |
| AGL(eBIC-eBIC) | 2.96 | 2.49 |
| Our Method | 3 | 2.18 |

*The simulation results are based on 100 replications with sample size $n = 150$. GL means group lasso method, AGL means adaptive group lasso method. The details of GL and AGL methods can be found in Lian (2012) and eBIC means extended Bayesian information criterion (Chen and Chen 2008).*

**6. Real data analysis.** We now apply the proposed method to analyse an environmental data set from Hong Kong. This data set was collected between January 1, 1994 and December 31, 1995. It is a collection of numbers of daily total hospital admissions for circulationary and respirationary problems, measurements of pollutants and other environmental factors in Hong Kong. The collected environmental factors are $SO_2$ (coded by $x_1$), $NO_2$ (coded by $x_2$), dust (coded by $x_3$), temperature (coded by $x_4$), change of temperature (coded by $x_5$), humidity (coded by $x_6$), and ozone (coded by $x_7$). What we are interested in is which environmental factors among the collected factors have significant effects on the number of daily total hospital admissions for circulationary and respirationary problems (coded by $y$), and whether the impacts of those factors vary over time (coded by $U$). As the

numbers of daily total hospital admissions are count data, it is natural to use Poisson regression model with varying coefficients, namely (5.1), to fit the data.

We apply the proposed group SCAD method to identify the significant variables and the nonzero constant coefficients, and estimate the functional or constant coefficients in the selected model. The kernel function used is still taken to be the Epanechnikov kernel, and the bandwidth is chosen to be $0.75[(\log d_n)/n]^{0.2}100\%$ of the range of the time. The tuning parameters $\lambda_1$, $\lambda_2$, $\lambda_4$ and $\lambda_4^*$ are selected by the data driven approach described in Section 4.2.

The selected model is

$$\mathsf{P}(y_i = k) = \frac{\xi_i^k}{k!} e^{-\xi_i}$$

with

$$\log(\xi_i) = a_0(U_i) + a_2(U_i)x_{i2} + a_4(U_i)x_{i4} + a_5(U_i)x_{i5} + a_6(U_i)x_{i6}.$$

This shows only variables $NO_2$, temperature, change of temperature, and humidity have effects on the number of daily total hospital admissions for circulationary and respirationary problems, and all of these variables have time-varying impacts. The estimates of the impacts of these variables are presented in Figure 1.

Figure 1 shows that $NO_2$ always has a positive impact on the daily number of total hospital admissions for circulationary and respirationary problems, and this impact is stronger in winter and spring than that in summer and autumn. This is in line with the finding in one World Health Organization report (WHO report, 2003) which shows some evidence that "long-term exposure to $NO_2$ at concentrations above 40–100 $\mu g/m^3$ may decrease lung function and increase the risk of respiratory symptoms". The nonlinear dynamic pattern of the impact of $NO_2$ also makes sense. This is because the main source of $NO_2$ pollution comes from the burning of coals and gasolines. In the winter and spring season, heating requirement will increase the amount of $NO_2$ pollution. This is evident from the plot of $NO_2$ in the data set. Furthermore, the fog and mist in winter and spring will also increase the chance that people expose to $NO_2$. Though $NO_2$ is toxic by inhalation, as its compound is acrid and easily detectable by smell at low concentrations, in most cases, the inhalation exposure to $NO_2$ can be generally avoided. However, when $NO_2$ is dissolved into the fog, this acid mist will be hard to

detect, and people may easily expose to this toxic acid mist for a long time without awareness.

Figure 1 also shows the change of temperature has a time-varying positive impact on the daily number of total hospital admissions for circulationary and respirationary problems. This coincides with the intuition that a sudden change of temperature would greatly increase the risk of catching cold, fever and other upper respirationary diseases. The impact of temperature is also time varying and mostly negative. It is stronger in autumn and spring than that in other seasons. This makes sense, indeed, colder autumn or spring would see more people catching circulationary or respirationary diseases.

The impact of humidity on the daily number of total hospital admissions for circulationary and respirationary problems is interesting and complicated. It does not seem to have any seasonal pattern. This is in line with the findings reported in the literature. Indeed, existing research (Strachan and Sanders, 1989; Schwartz, 1995; and Leon *et al*, 1996) agrees that humidity has a signifiant effect on daily hospital admissions for circulationary and respirationary problems in many different places. Strachan and Sanders (1989) study the childhood respiratory problems against the indoor air temperature and relative humidity. Through a randomly sampled questionnaire survey, and interview of $1,000$ children aged 7 about their living conditions and reported circulationary and respirationary problems, they show that the children living in damp (higher relative humidity level) bedrooms had significantly higher probability to catch day cough, night cough and chesty colds. Schwartz (1995) studies the short term fluctuations in air pollution and hospital admissions of the elderly for respiratory disease. According to their data set, the risk, measured by sample variance, of respiratory hospital admissions of people aged 65 or above is bigger in the cities with higher average humidity levels (measured by dew point). Leon *et al* (1996) study the effects of air pollution on daily hospital admissions for respiratory disease based on a data set collected in London between 1987-88 and 1991-92. They show that the relative humidity is more significant for the respiratory hospital admission numbers of children (0-14 years) and the elderly (65+ years). All of these suggest that there may be a strong relationship between humidity level and the risk for children and elderly people to catch circulationary or respirationary disease.

Furthermore, we would like to examine the prediction performance of the selected model and compare it with the full model with functional coefficients. Given either model, we begin with using the first 700 observations as the training set to estimate the conditional expectation of the response

variable of the 701st observation. Then we repeat this one-step forward prediction by enrolling one more observation into the training set at a time. Finally, we end with using the first 729 observations to predict the 730th observation. The prediction performance is measured by the mean relative prediction error (MRPE) defined as follows,

$$\text{MRPE} = \frac{1}{30}\sum_{i=701}^{730}\left|\frac{\hat{\xi}_i - y_i}{y_i}\right| \times 100\%,$$

where $\hat{\xi}_i$ is the estimator of the conditional expectation of the response variable at time $U_i$, $i = 701, \cdots, 730$. The MRPE of the model selected by our method is 18.7% while the MRPE of the full model with functional coefficients is 41.6%. Hence, we can see that the model selected by our method do have a better prediction accuracy than the full model.

**7. Acknowledgements.** The authors thank the Co-Editor, an Associate Editor and three referees for the helpful comments which greatly improved the former version of the paper.

# Appendix

In Appendix A, we give some regularity conditions which are needed to prove the asymptotic theory. In Appendices B and C of the supplemental material [Li *et al* (2015)], we provide the proofs of the main theoretical results and some auxiliary results, respectively.

## APPENDIX A: ASSUMPTIONS

Recall that

$$q_1(s,y) = \frac{\partial \ell\big[g^{-1}(s),y\big]}{\partial s}, \quad q_2(s,y) = \frac{\partial^2 \ell\big[g^{-1}(s),y\big]}{\partial s^2}$$

and define

$$\ddot{\mathcal{L}}_n(u) = \begin{bmatrix} \ddot{\mathcal{L}}_n(u,0) & \ddot{\mathcal{L}}_n(u,1) \\ \ddot{\mathcal{L}}_n(u,1) & \ddot{\mathcal{L}}_n(u,2) \end{bmatrix}$$

with

$$\ddot{\mathcal{L}}_n(u,l) = \frac{1}{n}\sum_{i=1}^{n} q_2\big[\sum_{j=1}^{d_n} a_j(U_i)x_{ij},\ y_i\big] X_i X_i^{\mathrm{T}} \left(\frac{U_i - u}{h}\right)^l K_h(U_i - u)$$

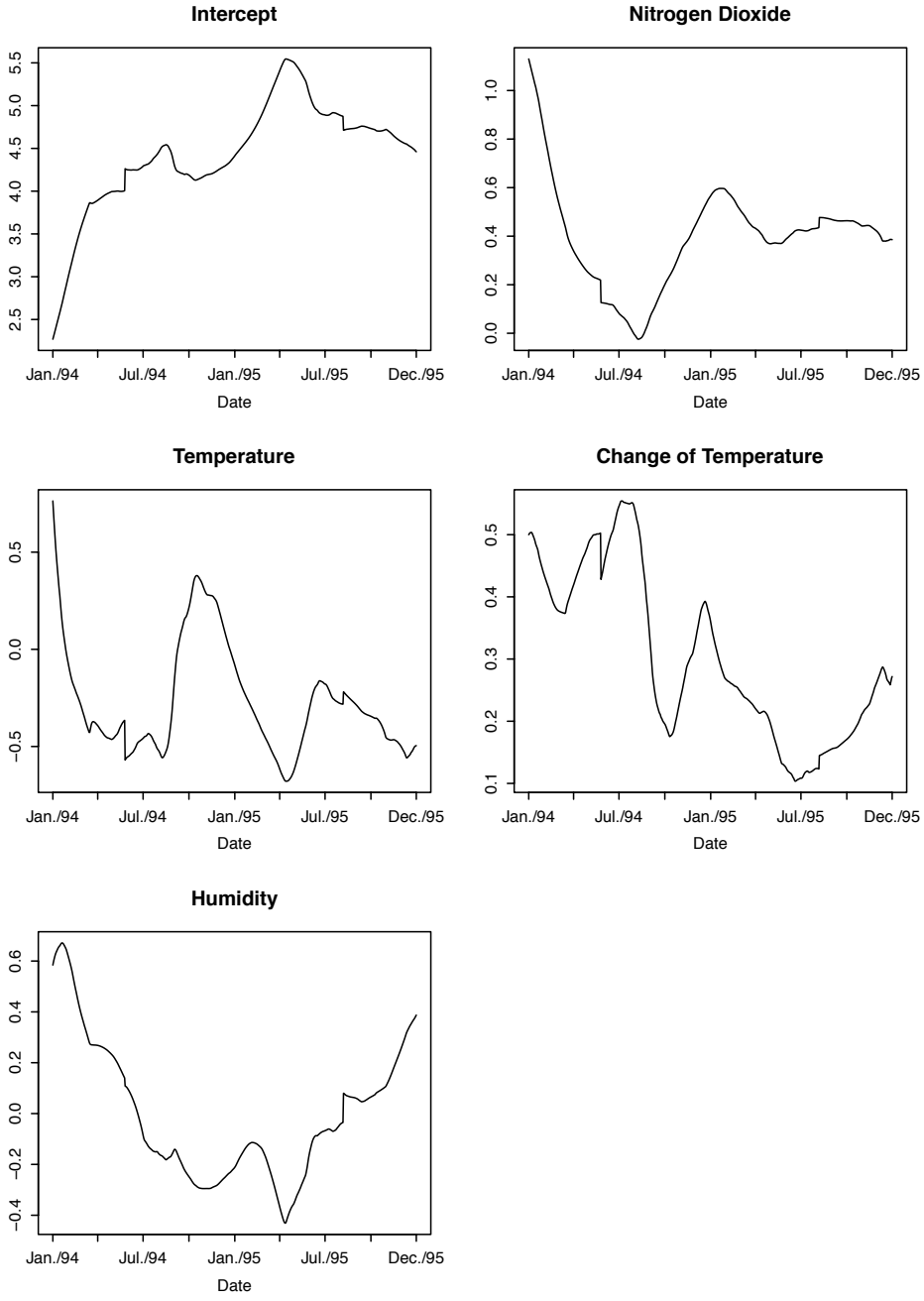


FIG 1. *Estimated curves of the functional coefficients in the selected model for the Hong Kong environment data.*

for $l = 0, 1, 2$. Define $b = \max\{\lambda_1/\lambda_2,\ \lambda_2/\lambda_1\} + \delta$ for any $\delta > 0$, where $\lambda_1$ and $\lambda_2$ are defined in (2.3), and let

$$\boldsymbol{\Omega}_0(b) = \Big\{\mathbf{v} = (v_{11}, \cdots, v_{1d_n}, v_{21}, \cdots, v_{2d_n})^{\mathrm{T}} :\ \|\mathbf{v}\| = 1,$$
$$\sum_{j=1}^{d_n}(|v_{1j}| + |v_{2j}|) \le 2(1+b)\sum_{j=1}^{s_{n2}}(|v_{1j}| + |v_{2j}|)\Big\}.$$

When $\lambda_1 \propto \lambda_2$ (see Assumption 5 or 5′ below), $b$ is bounded by a positive constant, and it becomes $1 + \delta$ which could be sufficiently close to 1 if we further assume that $\lambda_1 = \lambda_2$. To simplify the presentation, we denote

$$Q_{i1} = q_1\big[\sum_{j=1}^{d_n} a_j(U_i)x_{ij}, y_i\big],\quad Q_{i2} = q_2\big[\sum_{j=1}^{d_n} a_j(U_i)x_{ij}, y_i\big].$$

We next introduce some regularity conditions which are needed to establish the asymptotic theory for the proposed model selection and structure specification procedure. Some of the conditions might be not the weakest possible conditions.

*Assumption 1.* The kernel function $K(\cdot)$ is a continuous and symmetric probability density function with a compact support.

*Assumption 2.* *(i)* Let $\mathsf{E}\big(Q_{i1}\big|X_i, U_i\big) = 0$ a.s., and $\mathsf{E}\big(Q_{i1}^2\big|U_i = u\big)$ be continuous for $u \in [0,\ 1]$. Moreover, suppose that uniformly for $u \in [0, 1]$, either

(A.1)

$$\max_{1\le j\le d_n} \mathsf{E}\big[|Q_{i1}x_{ij}|^{m_0}\big|U_i = u\big] + \max_{1\le j,k\le d_n} \mathsf{E}\big[|Q_{i2}x_{ij}x_{ik}|^{m_0}\big|U_i = u\big] < \infty\ \ a.s.$$

for $m_0 > 2$, or

(A.2)

$$\max_{1\le j\le d_n} \mathsf{E}\big[|Q_{i1}x_{ij}|^{m}\big|U_i = u\big] + \max_{1\le j,k\le d_n} \mathsf{E}\big[|Q_{i2}x_{ij}x_{ik}|^{m}\big|U_i = u\big] \le \frac{M_0 m!}{2}\ \ a.s.$$

for all $m \ge 2$ and $0 < M_0 < \infty$.

*(ii)* Let $q_2(s, y) < 0$ for $s \in \mathbb{R}$ and $y$ in the range of the response variable. Furthermore, there exists an $M(X, U, y) > 0$ such that

$$\Big|q_2\big[r(X,U) + \delta_*, y\big] - q_2\big[r(X,U), y\big]\Big| \le M(X,U,y)|\delta_*|$$

with $r(X, U) = \sum_{j=1}^{d_n} a_j(U)x_j$, and uniformly for $u \in [0, 1]$ either

$$\max_{1\le j,k,l\le d_n} \mathsf{E}\big[|x_{ij}x_{ik}x_{il}M(X_i, U_i, y_i)|^{m_0}\big|U_i = u\big] < \infty\ \ a.s.$$

for $m_0 > 2$ if (A.1) is satisfied, or

$$\max_{1\le j,k,l\le d_n} \mathsf{E}\big[|x_{ij}x_{ik}x_{il}M(X_i,U_i,y_i)|^m|U_i=u\big] < \frac{M_1 m!}{2} \quad a.s.$$

for all $m \ge 2$ if (A.2) is satisfied, $0 < M_1 < \infty$.

*(iii)* There exist $0 < \rho_1 \le \rho_2 < \infty$ such that

$$\rho_1 \le \inf_{u\in[0,1]}\inf_{\mathbf{v}\in\mathbf{\Omega}_0(b)} \mathbf{v}^{\mathrm{T}}\big[-\ddot{\mathcal{L}}_n(u)\big]\mathbf{v} \le \sup_{u\in[0,1]}\sup_{\mathbf{v}\in\mathbf{\Omega}_0(b)} \mathbf{v}^{\mathrm{T}}\big[-\ddot{\mathcal{L}}_n(u)\big]\mathbf{v} \le \rho_2$$

with probability approaching one.

*Assumption 3.* The density function $f_U(\cdot)$ has a continuous second-order derivative. In addition, $f_U(u)$ is bounded away from zero and infinity when $u \in [0,\ 1]$.

*Assumption 4.* The functional coefficients, $a_j(\cdot)$, have continuous second-order derivatives for $j = 1, \cdots, d_n$.

*Assumption 5.* Let $d_n \propto n^{\tau_1}$ and $\frac{nh}{(nd_n^3)^{2/m_0}\log h^{-1}} \to \infty$, where $0 \le \tau_1 < \infty$ and $m_0$ is defined in (A.1). Moreover, the bandwidth $h$ and the tuning parameters $\lambda_1$ and $\lambda_2$ satisfy $h \propto n^{-\delta_1}$ with $0 < \delta_1 < 1$, $s_{n2}h^2 + \big(\frac{\log h^{-1}}{nh}\big)^{1/2} = o(\lambda_1)$, $\lambda_1 \propto \lambda_2$ and $s_{n2}\lambda_1^2 h^{-2} + s_{n2}^2\lambda_1 = o(1)$.

*Assumption 5′.* Let $d_n \propto \exp\big\{(nh)^{\tau_2}\big\}$ with $0 \le \tau_2 < 1-\tau_3$, $0 < \tau_3 < 1$. Furthermore, the bandwidth $h$ and the tuning parameters $\lambda_1$ and $\lambda_2$ satisfy $h \propto n^{-\delta_1}$ with $0 < \delta_1 < 1$, $s_{n2}h^2 + \big(\frac{\log h^{-1}}{nh}\big)^{\tau_3/2} = o(\lambda_1)$, $\lambda_1 \propto \lambda_2$ and $s_{n2}\lambda_1^2 h^{-2} + s_{n2}^2\lambda_1 = o(1)$.

*Assumption 6.* *(i)* Let $s_{n2}h^2 \propto (nh)^{-1/2}$, $\lambda_4 \sim \lambda_4^*$, $\lambda_4 = o\big(n^{1/2}\big)$, and

$$h^{-1/2}\big[(\log h^{-1})^{1/2} + s_{n2}^{1/2}(1+\lambda_1\sqrt{nh})\big] = o(\lambda_4). \tag{A.3}$$

*(ii)* There exists a positive constant $b_\diamond$ such that

$$\min_{1\le j\le s_{n2}} \|\boldsymbol{\alpha}_{j0}\| \ge b_\diamond n^{1/2}, \quad \min_{1\le j\le s_{n1}} D_j \ge b_\diamond n^{1/2} \tag{A.4}$$

with probability approaching one. Furthermore, uniformly for $k = s_{n2}+1, \cdots, d_n, d_n+s_{n1}+1, \cdots, 2d_n$,

$$\sup_{u\in[0,1]}\sup_{\|\mathbf{w}^o\|=1} \Big|\ddot{\mathcal{L}}_n(u|k)\mathbf{w}^o\Big| = O_P(1), \tag{A.5}$$

where $\ddot{\mathcal{L}}_n(u|k)$ is the $k$-th row of $\ddot{\mathcal{L}}_n(u)$, $\mathbf{w}^o = \big[(\mathbf{w}_1^o)^{\mathrm{T}}, (\mathbf{w}_2^o)^{\mathrm{T}}\big]^{\mathrm{T}}$, $\mathbf{w}_1^o$ and $\mathbf{w}_2^o$ are two $d_n$-dimensional column vectors, the last $d_n - s_{n2}$ elements of $\mathbf{w}_1^o$ and the last $d_n - s_{n1}$ elements of $\mathbf{w}_2^o$ are zeros.

**Remark A.1.** The above assumptions are mild and justifiable. Assumption 1 is a commonly-used condition on the kernel function and can be satisfied for the uniform kernel function and the Epanechnikov kernel function which is used in our numerical study. The compact support restriction on the kernel function is not essential and can be removed at the cost of more tedious proofs. Assumption 2 imposes some smoothness and moment conditions on $Q_{i1}$ and $Q_{i2}$, some of which are commonly used in local maximum likelihood estimation (c.f., Cai *et al*, 2000, Li and Liang, 2008). Two moment conditions (A.1) and (A.2) are imposed in Assumption 2(i), and they are used to handle the polynomially diverging dimension of the covariates (in Assumption 5) and the exponentially diverging dimension of the covariates (in Assumption 5′), respectively. Hence, as the dimension of the covariates increase from the polynomial order to the exponential order, the required moment condition would be stronger. In contrast, most of the existing literature such as Lian (2012) only considers the case of the stronger moment condition in (A.2), which may possibly limit the applicability of the model selection methodology. Assumption 2(iii) can be seen as the modified version of the so-called *restricted eigenvalue condition* introduced by Bickel *et al* (2009) for the parametric regression models. Assumptions 3 and 4 provide some smoothness conditions on the density function of $U$ and the functional coefficients $a_j(\cdot)$, which are not uncommon when the local linear approach is applied (c.f., Fan and Gijbels, 1996).

Assumption 5 imposes some restrictions on the bandwidth $h$ and the tuning parameters $\lambda_1$ and $\lambda_2$ when $d_n \propto n^{\tau_1}$, whereas Assumption 5′ imposes some conditions when $d_n \propto \exp\{(nh)^{\tau_2}\}$. They are crucial to derive the uniform convergence rates for the preliminary estimation in Proposition 3.1. Consider a special case when $s_{n2}$ is a fixed positive integer, we may choose $h \propto n^{-1/5}$ and $\lambda_1 = \lambda_2 \propto n^{-3/10}$. Then, the conditions in Assumption 5 would be satisfied when $m_0$ is sufficiently large, and those in Assumption 5′ would be satisfied when $3/4 < \tau_3 < 1$. Noting that $s_{n2}h^2 + \big(\frac{\log h^{-1}}{nh}\big)^{1/2} = o(\lambda_1)$ and $\lambda_1 \propto \lambda_2$ by Assumption 5, the influence by $h$ and $\lambda_2$ on the uniform convergence rate in (3.1) is dominated by that of $\lambda_1$. The regularity conditions in Assumption 6 is mainly used to prove the sparsity and oracle property for the proposed model selection procedure. By Assumption 5, we may show that the leading term of the left hand side of (A.3) is $\lambda_1\sqrt{ns_{n2}}$. Once again we consider the special case when $s_{n2}$ is a fixed positive integer and $h \propto n^{-1/5}$. Then, the conditions in Assumption 6 would be satisfied if we choose $\lambda_1 \propto n^{-3/10}$ and $\lambda_4 = \lambda_4^* \propto n^{\tau_4}$ with $1/5 < \tau_4 < 1/2$.

## SUPPLEMENTARY MATERIAL

**Supplement to "Model selection and structure specification in ultra-high dimensional generalised semi-varying coefficient models"**. We provide the detailed proofs of the main results stated in Section 3 as well as some technical lemmas which are useful in the proofs.

## REFERENCES


Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, **37**, 1705–1732.

Bradic, J., Fan, J. and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh-dimensional variable selection. *Journal of Royal Statistics Society Series B*, **73**, 325–349.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics, Springer.

Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, **95**, 888–902.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**, 759–771.

Cheng, M., Honda, T., Li, J. and Peng, H. (2014). Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. *The Annals of Statistics*, **42**, 1819–1849.

Cheng, M., Zhang, W. and Chen, L. (2009). Statistical estimation in generalized multiparameter likelihood models. *Journal of the American Statistical Association*, **104**, 1179–1191.

de Leon, A. P., Anderson, H. R., Bland, J. M., Strachan, D. P., and Bower, J. (1996). Effects of air pollution on daily hospital admissions for respiratory disease in London between 1987-88 and 1991-92. *Journal of Epidemiology and Community Health*, **50**(Suppl 1), s63-s70.

Efron, B., Hastie, T. J., Johnstone, I. and Tibshirani, R. J. (2004). Least angle regression (with discussion). *The Annals of Statistics*, **32**, 407–499.

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *Journal of the American Statistical Association*, **106**, 544-557.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & 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 Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B*, **70**, 849–911.

Fan, J. and Lv, J. (2011). Non-concave penalized likelihood with NP-Dimensionality. *IEEE: Information Theory*, **57**, 5467–5484.

Fan, J., Ma, Y. and Dai, W. (2014). Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association*, **109**, 1270–1284.

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research*, **10**, 2013–2038.

Fan, J., and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, **38**, 3567–3604.

Fan, J., Xue, L. and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics*, **42**, 819–849.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics*, **27**, 1491–1518.

Fan, J. and Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics*, **27**, 715–731.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society Series B*, **75**, 531–552.

Huang, J., Horowitz, J. L. and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics*, **36**, 587–613.

Huang, J., and Xie, H. (2007). Asymptotic oracle properties of SCAD-penalized least squares estimators. *Lecture Notes-Monograph Series*, 149–166.

Hunter, D. and Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics*, **33**, 1617–1642.

Janson, S. (1987). Maximal spacing in several dimensions. *The Annals of Probability*, **15**, 274–280.

Kai, B., Li, R. and Zou, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *The Annals of Statistics*, **39**, 305–332.

Li, D., Ke, Y. and Zhang, W. (2013). Model selection in generalised semi-varying coefficient models with diverging number of potential covariates. Working paper, Department of Mathematics, University of York.

Li, D., Ke, Y. and Zhang, W. (2015). Supplement to "Model selection and structure specification in ultra-high dimensional generalised semi-varying coefficient models".

Li, J. and Zhang, W. (2011). A semiparametric threshold model for censored longitudinal data analysis. *Journal of the American Statistical Association*, **106**, 685–696.

Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modelling. *The Annals of Statistics*, **36**, 261–286.

Lian, H. (2012). Variable selection for high-dimensional generalized varying-coefficient models. *Statistica Sinica*, **22**, 1563–1588.

Liu, J., Li, R. and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh dimensional covariates. *Journal of the American Statistical Association*, **109**, 266–274.

Schwartz, J. (1995). Short term fluctuations in air pollution and hospital admissions of the elderly for respiratory disease. *Thorax*, **50**(5), 531-538.

Song, R., Yi, F. and Zuo, H. (2012). On varying-coefficient independence screening for high-dimensional varying-coefficient models. Forthcoming in *Statistica Sinica*.

Strachan, D. P. and Sanders, C. H. (1989). Damp housing and childhood asthma; respiratory effects of indoor air temperature and relative humidity. *Journal of Epidemiology and Community Health*, **43**(1), 7-14.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B*, **58**, 267–288.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying-coefficient model. *Journal of the American Statistical Association*, **104**, 747–757.

Wang, L., Kai, B. and Li, R. (2009). Local rank inference for varying coefficient models. *Journal of American Statistical Association*, **104**, 1631–1645.

Wang, L. F., Li, H. Z. and Huang, J. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, **103**, 1556–1569.

Wei, F., Huang, J. and Li, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica*, **21**, 1515–1540.

World Health Organization (2003). Health aspects of air pollution with particulate matter, ozone, and nitrogen dioxide. *Rep. EUR/03/5042688*, Bonn.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, **68**, 49–67.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**, 894–942.

Zhang, C. H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics*, **36**, 1567–1594.

Zhang, W., Fan, J. and Sun, Y. (2009). A semiparametric model for cluster data. *The Annals of Statistics*, **37** , 2377–2408.

Zhang, W. and Peng, H. (2010). Simultaneous confidence band and hypothesis test in generalized varying-coefficient models. *Journal of Multivariate Analysis*, **101**, 1656–1680.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *The Annals of Statistics*, **36**, 1509–1566.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, **37**, 1773.

Department of Mathematics
The University of York
Heslington
York YO10 5DD
The United Kingtom
E-mail: degui.li@york.ac.uk

Department of Mathematics
The University of York
Heslington
York YO10 5DD
The United Kingtom
E-mail: yk612@york.ac.uk

Department of Mathematics
The University of York
Heslington
York YO10 5DD
The United Kingtom
E-mail: wenyang.zhang@york.ac.uk