Feature selection and structure specification in ultra-high dimensional semi-parametric model with an application in medical science

Yuan Ke

Ph.D.

UNIVERSITY OF YORK

MATHEMATICS

March 2015

Abstract

In this thesis, we consider the feature selection, model specification and estimation of the generalised semi-varying coefficient models (GSVCMs), where the number of potential covariates is allowed to diverge with the sample size. Based on the penalised likelihood approach and kernel smoothing method, we propose a penalised weighted least squares procedure to select the significant covariates, identify constant coefficients among the coefficients of the selected covariates, and estimate the functional or constant coefficients in GSVCMs. A computational algorithm is also proposed to implement the procedure. Our approach not only inherits many desirable statistical properties from the local maximum likelihood estimation and nonconcave penalised likelihood method, but also computationally attractive thanks to the proposed computational algorithm. Under some mild conditions, we establish the theoretical properties for the proposed procedure such as sparsity, oracle property and the uniform convergence rates of the proposed estimators. We also provide simulation studies to show the proposed procedure works very well when the sample size is finite. We then use the proposed procedure to analyse a real environmental data set, which leads to some interesting findings. Finally, we establish a classification method and show it can be used to improve predictive modelling for classify the patients with early inflammatory arthritis at baseline into different risk groups in future disease progression.

Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I want to express my heartily gratitude to my supervisor, Professor Wenyang Zhang, for his continuous support to my Ph.D study and research, for his enthusiasm, motivation, deep thoughts, and immense knowledge. I sincerely appreciate all the time and effort he spent on training my research abilities. The independent, creative and critical thinking he taught me will be the treasure of my life.

I also want to thank Dr. Degui Li and Dr. Stephen Connor for being my TAP panel members and giving me many good advice on my study and research. I want to thank Dr. Degui Li for helpful discussions. I would like to acknowledge the Department of Mathematics for all the support.

Furthermore, cheers to my dear friends and colleges: Dr. Hongjia Yan, Mr. John Box and Mr. Xiang Li for all the support, interesting discussions, and beers. I want to thank Mr. John Box for his comments to improve the thesis.

Finally, I want to say thanks to my family: my father Wenjin Ke, my mother Xiaojuan Zhou and my wife Yan Huo. Their love, support and understanding is always the light guiding me through thick and thin.

Author's Declaration

Chapter 2.1 is an introduction to the framework of local polynomial modelling, which is mostly from the book: Fan, J. and Gijbels, I. (1996). *Local Polynomial modelling and Its applications*.

Chapter 2.2 is a review of the penalised likelihood method, which is mainly based on (but not limited to) the following literature: Tibshirani (1996), Fan and Li (2001), Zou and Li (2008) and Fan and Lv (2010).

Chapter 8 is mainly based on my paper:
"A semi-varying coefficient multinomial logistic regression for prognostic classification with application to stratified medicine", joint with Dr. Bo Fu (Centre for Biostatistics & Arthritis Research UK Epidemiology Unit, The University of Manchester, United Kingdom) and Professor Wenyang Zhang.

The rest chapters are based on the following two papers of mine:
"Model selection in generalised semi-varying coefficient models with diverging number of potential covariates", joint with Dr. Degui Li and Professor Wenyang Zhang; and
"Model selection and structure specification in ultra-high dimensional generalised semi-varying coefficient models", joint with Dr. Degui Li and Professor Wenyang Zhang.

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, university. All sources are acknowledged as references.

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 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. A popular penalty function is the L_1 penalty, which leads to the LASSO (Tibshirani 1996). Efron *et al* (2004) developed an efficient algorithm to solve the entire solution path of the LASSO. Yuan and Lin (2006) extended the LASSO to group selection and proposed the group LASSO. Whilst the LASSO comes with many nice properties, it is biased. Zou (2006) proposed the adaptive LASSO to fix the inconsistency problem of the LASSO. Fan and Li (2001) argued that nonconcave penalty function would serve better than concave one, such as L_1, in model selection. They built an unified nonconcave penalised likelihood framework, and proposed a penalty function termed SCAD as an example of their framework. They showed the SCAD enjoys the properties of unbiasedness, sparsity and continuity. Further works on the nonconcave penalised likelihood method, such as its application in survival models, varying coefficient

models, can be found in Fan and Li (2002), Li and Liang (2008). In the implementation of the nonconcave penalised likelihood method, the optimisation of the penalised likelihood function is challenging. Hunter and Li (2005) proposed an MM algorithm to meet the challenge, and proved the convergence of the algorithm. Zou and Li (2008) applied the local linear approximation to a nonconcave penalty function, and showed the algorithm for LASSO can be used to compute the solution of nonconcave penalised least squares.

In high-dimensional data analysis, it is often the case that the number of potential covariates grow beyond the sample size. For parametric models, there is literature addressing this problem, see Huang *et al* (2008), Zhang and Huang (2008), Zou and Zhang (2009), Huang and Xie (2007). Furthermore, some literature explored the ultra-high dimensional cases, allowing the number of potential covariates diverge with certain exponential rate, see Fan and Lv (2008), Fan *et al* (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 applications. They often lead to model misspecification, which would result in inconsistent estimates and incorrect conclusions being drawn from the data analysed. In this thesis, we relax this linear restriction and use functional coefficients to describe the relationship between response variables and covariates. Varying coefficient models, as an useful generalisation of linear models, have played an important role in the analysis of 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).

Like any other family of models, model selection in the varying coefficient models is of great interest and has been extensively studied in the literature. For instance, Wang *et al* (2008) and Wang and Xia (2009) use group penalisation to select the significant variables in varying coefficient models when the number of potential covariates is fixed. More recently, for 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. Lian (2012) considers variable selection in generalised varying coefficient models whilst allowing that the number of covariates to diverge with the sample size.

Unlike the literature, in this thesis, the model selection for the proposed varying coefficient models has two aspects: (i) variable selection; and (ii) identification of the constant coefficients. We remark that variable selection is equivalent to identifying the zero functional coefficients and that 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. With this in mind, we call the proposed model selection procedure feature selection and model specification.

Suppose we have a response variable y, covariate U, and potential covariates $x_1, \cdots, x_{d_n}$, where d_n depends on sample size n, and $d_n \to$

∞ when $n \to \infty$. Let $X = (x_1, \cdots, x_{d_n})^{\mathrm{T}}$,

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

be the conditional expectation of y given (U, X^{T}). In this thesis, we define the density function of a discrete random variable as its probability mass function. We assume the log conditional density function of y given X and U is

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

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 and to be estimated, and $C_1(\boldsymbol{\phi}_1) > 0$, $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ are unknown nuisance parameters.

The family of models in (1.1) is a natural extension of generalised linear models by allowing the coefficients to vary with the index variable U and for some functional coefficients to possibly be constant. Hence we term (1.1) *generalised semi-varying coefficient models* (GSVCMs). The family of GSVCMs is not only a mathematical generalisation but also stimulated by the demands in real applications. In the following part, we give two brief examples to illustrate the usage of GSVCMs in practice. The analysis of these examples can be considered as some future applications of the methods proposed in this thesis.

Example 1.1. *Estimation of the perk time in magnetic resonance imaging (MRI) scan.*

MRI scan is a radiological imaging technique which is widely used in many medical studies and clinical fields. For some MRI scans, the patients need to have an injection of contrast dye to make certain tissues and blood vessels show up more clearly and in greater detail. After the injection, the contrast dye will spread and be absorbed by the body gradually and its concentration in blood will increase first and then decrease. So as to make a clear enough image, MRI scans need to be taken at the perk time, i.e. the time window that the contrast dye concentration in blood is at a high level. In order to guarantee a successful scan, the hospitals used to use a relatively large dose of contrast dye to gain a long enough perk time. However, the injection of contrast dye may cause many severe side-effects such as nausea, vomiting, urticaria, anaphylaxis and so on. It is always desirable to keep the injection dose of contrast dye at a low level. Hence an accurate estimation of the perk time will be the key to reduce the injection dose of contrast dye. The traditional way to estimate the perk time is to establish a Poisson regression model between the perk time (response variable) and some covariates like gender, height, weight, blood pressure, average heart rate, injection dose, injection rate, and so on. In practice, this model does not work well as the impacts of the covariates will depend on the age of the patients in a complicated way. As an alternative, it will be natural to consider a semi-varying coefficient Poisson regression model and take the age of the patients as the covariate U.

Example 1.2. *Prediction of stock market movements using Internet data.*

Nowadays, the Internet has become an indispensable part of society. Compared with traditional media sources like newspapers and television, one big advantage of the Internet is how powerful search engines allow people to obtain the information they want to know at any time. Hence the records of search engines can be used as important data to analyse their users including their behaviour in the stock market. In this example, one can analyse a stock index by making use of the search volume of various terms. One flexible model assumption is to assume the stock index is a semi-varying coefficient model of search volumes and the impact of these search volumes may vary with some index covariate U. For example, the index covariate U can be chosen as time, income, location or some other covariate. Furthermore, one can allow the number of potential covariates to diverge with the sample size and do model selection by the methods proposed in this thesis.

In this thesis, we will investigate feature selection and model specification procedure of GSVCMs under both diverging and ultra-high dimensionality. For the diverging dimensionality case, the methodology we are going to use is based on kernel smoothing, penalised likelihood estimation and group selection idea. We first obtain preliminary estimators of the functional coefficients using local linear approximation and log-likelihood estimation. Then, based on the preliminary estimators, we propose a penalised weighted least squares procedure with group selection penalty to select significant covariates, identify constant coefficients and estimate functional or constant coefficients. For the ultra-high dimensionality case, we first propose a penalised likelihood method with LASSO penalty function to obtain preliminary esti-

mators of the functional coefficients, which are proved to be uniformly consistent. The uniform convergence rate for preliminary penalised semi-parametric estimators relies on the number of non-zero functional coefficients and the tuning parameters associated in the penalty terms. Then, we use the preliminary estimators of the functional coefficients in the quadratic approximation for the local log-likelihood function, and the construction of the adaptive group LASSO penalty and the adaptive SCAD penalty. We 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. Compared with the preliminary estimators, the final estimators enjoy some nice statistical properties such as sparsity and oracle property.

For both cases, the developed feature selection and model specification approaches inherit many desirable statistical properties from both the local maximum likelihood estimation and non-concave penalised likelihood method. Under some regularity conditions, we establish some asymptotic properties for the proposed feature selection, model specification and estimation procedures such as the sparsity and oracle property. In order to implement our methods in practical applications, we further develop novel iterative computational algorithms to do the maximisation involved in the estimation procedure when the group/adaptive SCAD or LASSO penalty is used. The SCAD penalty has many advantages and is widely used in shrinkage method. The common approach to implement the SCAD penalty in shrink-

age method for varying coefficient models consists of two steps: (i) approximate SCAD with an L_1 penalty locally using local linear approximation; (ii) apply the quadratic approximation to deal with the L_1 penalty. In this thesis, we do not go down that route. Making use of the structure of the SCAD penalty, we propose a different algorithm to implement our method. Furthermore, our novel iterative computational algorithms have a "double check" mechanism which works as follows: If after an iteration a covariate is identified as insignificant or with constant coefficient, it still has a chance to be re-selected into the model or identified as with functional coefficient in the following iteration. Thanks to this "double check" mechanism, our methods have good model selection performance and are not very sensitive to the choice of the preliminary estimators. Our simulation results show that both the adaptive group LASSO and the adaptive SCAD methods perform reasonably well, with the latter giving slightly better performance. The method developed in this thesis outperforms those in Wang and Xia (2009), and Lian (2012).

In this thesis, we also establish a multicategory classification method based on semi-parametric predictive modelling. Our multicategory classification method is based on a semi-varying coefficient multinomial logistic regression model and contains three steps: (i) feature selection and model specification; (ii) coefficient estimation; and (iii) classification. We conduct simulation studies to assess our method's performance and the results show that its correct classification rates compare well with the oracle one which is based on the true model and true coefficients under different scenarios. We illustrate the use of

our method by applying it to classify the patients with early inflammatory arthritis at baseline into different risk groups in future disease progression and use a leave-one-out cross-validation method to assess its correct classification rate.

The rest of the thesis is organised as follows. We begin in Chapter 2 with a literature review on local polynomial modelling and penalised likelihood method. Chapter 3 describes the proposed feature selection and model specification procedures. Chapter 4 gives the asymptotic properties of the proposed feature selection and model specification procedures. Chapter 5 provides computational algorithms to implement the developed methods. Chapter 6 discusses how to select the tuning parameters. In Chapter 7.1, the performance of the proposed feature selection, model specification and estimation procedures and algorithms are illustrated by some simulation studies. We also compare the finite sample performance of our method with some existing ones. In Chapter 7.2, we apply the generalised semi-varying coefficient models together with the proposed feature selection, model 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 circulatory and respiratory problems in Hong Kong. In Chapter 8, we establish a multicategory classification method based on a semi-varying coefficient multinomial logistic regression model and analyse a prognostic classification problem in medical science. In the end, the regularity conditions, the proofs of the main theoretical results and some auxiliary results for diverging and ultra-high dimensional

GSVCMs are provided in Chapter 9 and Chapter 10, respectively.

2 Literature review

2.1 Framework of local polynomial modelling

We first review the framework of local polynomial regression. In order to get an insight into this technique, we start with the theoretical basis of applying local polynomial approximation to an *i.i.d.* bivariate data sample $(X_1, Y_1) \cdots (X_n, Y_n)$ from the population (X, Y). We assume the data are generated from the model

$$Y = m(X) + \sigma(X)\varepsilon \tag{2.1}$$

where the error ε is independent of X, $E(\varepsilon) = 0$, and $\mathrm{Var}(\varepsilon) = 1$. Our goal is to estimate the regression function $m(x_0) = E(Y|X = x_0)$ and its derivatives $\dot{m}(x_0), \ddot{m}(x_0), \cdots, m^{(p)}(x_0)$. It is assumed that the $(p+1)$th derivative of $m(x)$ at the point x_0 exists.

Through a Taylor expansion for the unknown regression function $m(x)$ in a neighbourhood of x_0, we can approximate it by a local polynomial as

$$\begin{aligned} m(x) \approx & m(x_0) + \dot{m}(x_0)(x - x_0) + \frac{\ddot{m}(x_0)}{2!}(x - x_0)^2 \\ & + \cdots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p. \end{aligned} \tag{2.2}$$

We can fit this polynomial by treating $\dfrac{m^{(j)}(x_0)}{j!} = \beta_j$ for $j = 0, 1, \cdots, p$, and solve them by minimizing the following weighted least squares regression:

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j \left(X_i - x_0 \right)^j \right\}^2 K_h \left(X_i - x_0 \right), \tag{2.3}$$

where h is a bandwidth, and $K_h(\cdot) = K(\cdot/h)/h$ is a kernel function allocating weights to each observation. Once we minimized this weighted least squares problem with respect to β_j and denote the minimizer by $\hat{\beta}_j$, $j = 0, 1, \cdots, p$, we can estimate the unknown function $m(x)$ and its derivatives by $\hat{m}^{(\nu)}(x_0) = \nu! \hat{\beta}_\nu$, $\nu = 0, 1, \cdots, p$.

Also, following the notations in Fan and Gijbels (1996), we can rewrite the weighted least squares problem in (2.3) in matrix form,

$$\min_{\beta} \left(\mathbf{y} - \mathbf{X}\beta \right)^T \mathbf{W} \left(\mathbf{y} - \mathbf{X}\beta \right), \tag{2.4}$$

where the design matrix $\mathbf{X}$, $\mathbf{y}$ and β are as follows:

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x_0) & \cdots & (X_n - x_0)^p \end{pmatrix},$$

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad , \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix},$$

and $\mathbf{W}$ is the $n \times n$ diagonal matrix of weights:

$$\mathbf{W} = \operatorname{diag} \left\{ K_h(X_i - x_0) \right\}.$$

It is easy to see the solution of (2.4) is given by

$$\hat{\beta} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}, \tag{2.5}$$

with $\hat{\beta} = (\hat{\beta}_0, \cdots, \hat{\beta}_p)^T$.

Recall that, the conditional expectation of $\mathbf{y}$ given $\mathbf{X}$ is $m(x_0) = E(Y|X = x_0)$. It is easy to see the conditional expectation and variance of $\hat{\beta}$ from (2.5):

$$\begin{aligned} E(\hat{\beta}|\mathbb{X}) &= \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{m} \\ &= \beta + (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{r} \\ \mathrm{Var}(\hat{\beta}|\mathbb{X}) &= \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\Sigma\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}, \end{aligned} \tag{2.6}$$

where $\mathbf{r} = \mathbf{m} - \mathbf{X}\beta$ is the residual vector, and

$$\Sigma = \mathrm{diag}\left\{K_h^2\left(X_i - x_0\right)\sigma^2(X_i)\right\}.$$

Although the exact conditional bias and variance of $\hat{\beta}$ have nice and simple closed forms, we can not directly use them since they involve unknown quantities, like the residual $\mathbf{r}$ and the diagonal matrix Σ. One way to solve this problem is to find the estimators of $\hat{\mathbf{r}}$ and $\hat{\Sigma}$, and plug them in equation (2.6). Another way, studied by Ruppert and Wand (1994), approximating the conditional bias and variance by their first order asymptotic expansions. Before illustrating their results in the following theorem, we would like to introduce some notations first. We denote the moments of K and K^2 by $\mu_j = \int u^j K(u)du$ and

$\nu_j = \int u^j K^2(u)du$ respectively. Let $e_{\nu+1} = (0, \cdots, 0, 1, 0, \cdots, 0)^T$ be the indicator vector with 1 on the $(\nu+1)^{th}$ position and 0 elsewhere. Also let

$$S = (u_{j+l})_{0\le j,l\le p}, \quad \tilde{S} = (u_{j+l+1})_{0\le j,l\le p}, \quad S^* = (u_{j+l})_{0\le j,l\le p},$$
$$c_p = (\mu_{p+1}, \cdots, \mu_{2p+1})^T, \quad \text{and} \quad \tilde{c}_p = (\mu_{p+2}, \cdots, \mu_{2p+2})^T.$$

In addition, we denote the conditional variance of Y given $X = x_0$ by $\sigma^2(x_0)$ and the marginal density of X by $f(\cdot)$. We now have the following theorem in Chapter 3.2 in Fan and Gijbels (1996).

Theorem 2.1. *Assume that $f(x_0) > 0$ and that $f(\cdot)$, $m^{(p+1)}(\cdot)$ and $\sigma^2(\cdot)$ are continuous in a neighbourhood of x_0. Further assume that $h \to 0$ and $nh \to \infty$. Then the asymptotic conditional variance of $\hat{m}_\nu(x_0)$ is given by*

$$Var(\hat{m}_\nu(x_0)|\mathbb{X}) = e_{\nu+1}^T S^{-1} S^* S^{-1} e_{v+1} \frac{\nu!^2 \sigma^2(x_0)}{f(x_0) n h^{1+2\nu}} + o_P\left(\frac{1}{nh^{1+2\nu}}\right). \tag{2.7}$$

The asymptotic conditional bias for $p - \nu$ odd is given by

$$Bias\{\hat{m}_\nu(x_0)|\mathbb{X}\} = e_{\nu+1}^T S^{-1} c_p \frac{\nu!}{(p+1)!} m^{(p+1)}(x_0) h^{p+1-\nu} + o_P(h^{p+1-\nu}). \tag{2.8}$$

Further, for $p - \nu$ even the asymptotic conditional bias is

$$Bias\{\hat{m}_\nu(x_0)|\mathbb{X}\} = e_{\nu+1}^T S^{-1} \tilde{c}_p \frac{\nu!}{(p+1)!} \{m^{(p+2)}(x_0) + (p+2) m^{(p+1)}(x_0) \frac{\dot{f}(x_0)}{f(x_0)}\} h^{p+2-\nu} + o_P(h^{p+2-\nu}), \tag{2.9}$$

provided that $\dot{f}(\cdot)$ *and* $m^{(p+2)}(\cdot)$ *are continuous in a neighbourhood of* x_0 *and* $nh^3 \to \infty$.

Now let us analyse the results showed in the above theorem. Suppose we fit an order p local polynomial to estimate the νth order derivative $m^{(\nu)}(x_0)$. The asymptotic bias of this fit has the order $h^{p+1-\nu}$ (for $p-\nu$ odd) or $h^{p+2-\nu}$ (for $p-\nu$ even). So, the bias order will decrease while p increases. This means fitting a higher order local polynomial will effectively decrease the asymptotic conditional bias. Does this suggest we should always apply a high order local polynomial regression? According to the conclusion in classical multivariate regression, an increase of approximate terms will result in an increase of the variance part. To see this from an intuitive point of view, when we choose a large p, each unknown term gets less "information" to estimate, and hence the approximation variability increases. So there exists a trade off between bias and variance associated with picking up a proper model order. The order of the asymptotic conditional variance according to (2.7) is $n^{-1}h^{-(1+2\nu)}$, which is not a function of order p. However, the order p does affect the constant term of the asymptotic conditional variance in a complicated way. The detailed discussion about this can be find in Ruppert and Wand (1994) and in Fan and Gijbels (1996) . An interesting and useful result of this discussion is that the asymptotic variance will not increase when moving from an $p-\nu$ even order to its consecutive odd order. The variability will only increase when the model moves from a $p-\nu$ odd order to its consecutive even order. In other words, when we estimate the regression function ($\nu = 0$), a $2p+1$ order fit, compared with a $2p$ order fit, introduces an extra pa-

rameter to reduce the bias, without paying price on the variance side. This result suggests that the odd order fits are preferable, i.e. local linear outperforms local constant, and local cubic outperforms local quadratic.

The choice of bandwidth is another important issue in local polynomial modelling. The bandwidth h will determine the size of the local neighbourhood of the polynomial fit, and its value will greatly affect the approximation result. For example, a bandwidth $h = 0$ corresponds to interpolating the data and choosing the most complicated model. On the contrary, $h = \infty$ leads to the simplest model – fitting a 'global' polynomial. Thus the bandwidth h, which runs from 0 to ∞, will play a role of controlling the model complexity. We can derive the asymptotically optimal bandwidth by minimizing the asymptotically conditional mean squared error with respect to the bandwidth h. However, the optimal bandwidth is not directly applicable in practice since it contains unknown quantities. Here we review a data-driven procedure to select a constant bandwidth in local polynomial fitting introduced in Fan and Gijbels (1996). The idea of this procedure are formed by three steps. In the first step, we derive good estimators of the bias and variance not fully relying on their asymptotic expressions. In the second step, we establish the Residual Squares Criterion (RSC) and obtain an optimal bandwidth estimator by minimizing the RSC. In the third step, we use the bandwidth estimator obtained in the second step as a pilot estimator and select the bandwidth which minimizes the estimated integrated mean squared error. This procedure is called the refined bandwidth selector.

Now we briefly introduce the main results about the refined bandwidth selector. The detailed results can be found in Chapter 4 of Fan and Gijbels (1996). Fan and Gijbels (1996) defined the Residual Squares Criterion (RSC) as the follows:

$$RSC(x_0; h) = \hat{\sigma}^2(x_0) \{1 + (p+1)V\}, \tag{2.10}$$

where $\hat{\sigma}^2(\cdot)$ is the estimator of the unknown variance $\sigma^2(\cdot)$, and V is the first diagonal element of the matrix

$$\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{W}^2\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}.$$

The estimator $\hat{\sigma}^2(\cdot)$ can be obtained by the normalized weighted residual sum of squares through a p-th order local polynomial approximation. Here we omit the corresponding technical details as they are not of interest in this thesis.

Then the optimal bandwidth can be estimated through the minimizer of the asymptotic expectation of the RSC statistic. The asymptotic expectation for the RSC statistic is given in the following theorem established by Fan and Gijbels (1995).

Theorem 2.2. *Suppose that $\sigma^2(x) = \sigma^2(x_0)$ in a neighbourhood of x_0. If $h_n \to 0$ and $nh_n \to \infty$, then*

$$\begin{aligned} E\{RSC(x_0; h_n)|\mathbb{X}\} =& \sigma^2(x_0) + C_p\beta_{p+1}^2 h_n^{2p+2} + (p+1)a_0\frac{\sigma^2(x_0)}{nh_n f(x_0)} \\ &+ o_P\left\{h_n^{2p+2} + (nh_n)^{-1}\right\}, \end{aligned}$$

where $C_p = \mu_{2p+2} - c_p^T S^{-1} c_p$ with $c_p = (\mu_{p+1}, \cdots, \mu_{2p+1})^T$, and $a_0 =$

$\int K_0^*(t)dt$ *which is the first diagonal element of the matrix* $S^{-1}S^*S^{-1}$.

This theorem reveals that the approximation of the minimizer of $E\{RSC(x_0;h_n)|\mathbb{X}\}$ with respect to h is

$$h_o(x_0) = \left\{ \frac{a_0\sigma^2(x_0)}{2C_p\beta_{p+1}^2 nf(x_0)} \right\}^{1/(2p+3)}. \tag{2.11}$$

And we have the simple relationship between h_o and the optimal bandwidth h_{opt} as

$$h_{opt}(x_0) = adj_{\nu,p}h_o(x_0), \tag{2.12}$$

where

$$adj_{\nu,p} = \left[\frac{(2\nu+1)C_p \int K_\nu^{*2}(t)dt}{(p+1-\nu)\left\{\int t^{p+1}K_\nu^*(t)dt\right\}^2 \int K_0^*(t)dt} \right]^{1/(2p+3)}.$$

As the adjusting constants $adj_{\nu,p}$ only depend on the kernel function, we find a statistic of which the minimizer leads to an estimator of h_{opt} and does not depend on unknown quantities.

Justified by the simulations in Fan and Gijbels (1995), the RSC bandwidth selection method gives good estimation of the optimal bandwidth, however the visual impression of the convergence rate of the estimator is not great for some cases. To improve the convergence rate of the bandwidth estimator, a refined bandwidth selector, introduced by Fan and Gijbels (1996), is as following:

> **Pilot estimation.** According to the RSC bandwidth selection method introduced in (2.11) and (2.12), get a pilot bandwidth estimator $h^* = \hat{h}_{p+1,p+2}^{RSC}$ by fitting a polynomial of order $p+2$.

Then use this pilot bandwidth h^* to obtain the estimates $\hat{\beta}_{p+1}$, $\hat{\beta}_{p+2}$ and $\hat{\sigma}^2(x_0)$.

Bandwidth selection. Find the minimizer of the following estimated integrated mean squared error:

$$\hat{h}^R_{\nu,p} = \arg\min_h \int_{[a,b]} \widehat{MSE}_{\nu,p}(y;h)dy, \tag{2.13}$$

where the detailed form of $\widehat{MSE}_{\nu,p}(y;h)$ is given in Chapter 4.3 in Fan and Gijbels (1996). Then use this bandwidth estimator obtained by the above refined bandwidth selector to fit a polynomial of order p.

According to the simulation results in Fan and Gijbels (1996), this refined bandwidth selector has higher relative rate of convergence than the RSC bandwidth selector. Fan and Gijbels (1996) also provide some simulations to show that a local polynomial model works well with the refined bandwidth selector and can adapt neatly to spatially inhomogeneous curves. This means even in the cases for which the curves show many alterations, and are very irregular, this data-driven methodology performs quite well.

2.2 Penalised likelihood method

In practice, a large number of predictors may be treated as candidates of true variables. So the initial step of high dimensional modelling is usually to select a proper subset from the large pool of candidates. We call this procedure the model selection or variable selection procedure. There are some variable selection techniques, which are practically useful in classical multivariate regression, like hypothesis testing, AIC and BIC coupled with computational algorithms such as, forward/backward search and stepwise deletion. However, problems occur when directly apply them into high dimensional model selection. There are several drawbacks of these methods. The first one is when dealing with high dimensional data, the number of stages involved in the variable selection procedure may be huge, and hence the stochastic errors inherited in each stage may accumulate and become very large. Another severe one is the theoretical properties of these classical variable selection methods are hard to derive under high dimensional model assumptions, i.e. "they lack of stability" as analysed in Breiman (1996). The third drawback is that the traditional best subset selection methods are usually computationally expansive. For example, the AIC or the BIC is often impractical for p dimensional data when p is large, since it would involve comparing 2^p models. One good alternative is the penalised likelihood approach. This approach attempts to automatically and simultaneously select significant covariates and estimate their coefficients via the combination of the likelihood function and a penalty function.

We will briefly review the theory of penalised least squares and the

SCAD penalty through the following linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x_n})^T$ is an $n \times p$ design matrix of covariates, $\mathbf{y} = (y_1, \cdots y_n)^T$ is an $n \times 1$ vector of the response variable, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ is an $n \times 1$ noise vector. Furthermore, for simplicity (but not necessary), we assume $\mathbf{X}$ has orthonormal columns, i.e. $\mathbf{X}^{\mathrm{T}}\mathbf{X} = nI_p$.

We define the penalised least squares (PLS) problem as

$$\min_{\boldsymbol{\beta} \in R^p} \left\{ \frac{1}{2n} \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \right\}, \tag{2.14}$$

where $\|\mathbf{u}\|^2 = \mathbf{u^T u}$, $p_\lambda(\cdot)$ is a penalty function, and $\lambda \geq 0$ is a tuning parameter. The first terms of (2.14) represents the goodness of fit while the second term represents the penalty for model complexity. Thus the minimizer of (2.14) can be also understood as a trade off between bias and variance.

By some calculations, the minimization problem in equation (2.14) can be transformed as:

$$\min_{\boldsymbol{\beta} \in R^p} \left\{ \frac{1}{2n} \left\| \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \right\|^2 + \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right\|^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \right\}, \tag{2.15}$$

where $\hat{\boldsymbol{\beta}} = n^{-1}\mathbf{X}^T\mathbf{y}$ is the ordinary least squares estimator. Therefore, minimizing (2.15) becomes a component-wise univariate PLS problem:

$$\hat{\theta}(z) = \arg\min_{\theta \in R} \left\{ \frac{1}{2}(z - \theta)^2 + p_\lambda(|\theta|) \right\}. \tag{2.16}$$

Now an important question is how to choose the penalty function. Antoniadis and Fan (2001) argues that a good PLS estimator $\hat{\theta}(z)$ should present three properties: sparsity, approximate unbiasedness and continuity. So Fan an Li (2001) pointed out an ideal choice of penalty function should be singular around the origin to produce sparsity, be upper bounded by a positive constant to produce approximate unbiasedness and satisfy $\arg\min_{t\geq 0}\{t + \dot{p}_\lambda(t)\} = 0$ to guarantee continuity.

As we know L_p penalties are widely used in penalised likelihood estimation. Some famous members of this penalty family are the L_2 penalty (Hoerl and Kennard, 1970), the L_1 penalty (LASSO) (Tibshirani 1996), or a combination of the two (Zou and Hastie, 2005). However all the L_p penalties can not satisfy all three aforementioned properties at the same time. For example, the concave L_p penalty with $0 \leq p < 1$ does not meet the continuity condition, the convex L_p penalties with $p > 1$ does not enjoy sparsity and the widely used convex L_1 penalty does not satisfy the approximate unbiasedness condition. For this reason, there is a need to find some penalty functions which satisfy these three properties simultaneously. One successful attempt, introduced by Fan (1997) and Fan and Li (2001), is the *smoothly clipped absolute deviation* (SCAD) penalty. The SCAD penalty is defined through its derivative as

$$\dot{p}_\lambda(t) = \lambda \left\{ I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right\} \quad \text{for some } a > 2, \quad (2.17)$$

where $p_\lambda(0) = 0$, and a is suggested to be 3.7 according to a Bayesian argument.

Furthermore, under some conditions, Fan and Li (2001) showed the resulting estimator of SCAD penalty enjoys Oracle property. In other words the estimator obtained by SCAD penalty works as well as the oracle estimator. Here, the oracle estimator means the estimator obtained when the correct sub-model were known.

Though the SCAD penalty enjoys many good properties, optimization of the penalised likelihood function with a non-convex penalty function is challenging. To solve this problem, Fan and Li (2001) proposed the local quadratic approximation (LQA) algorithm for non-concave penalty case. Using a Taylor expansion, and given an initial value $\beta^* = (\beta_1^*, \cdots, \beta_p^*)^T$, the penalty function p_λ can be locally approximated by a quadratic function as

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^*|) + \frac{1}{2}\frac{\dot{p}_\lambda(|\beta_j^*|)}{|\beta_j^*|}[\beta_j^2 - (\beta_j^*)^2], \qquad for\, \beta_j \approx \beta_j^*. \quad (2.18)$$

With this quadratic approximation, the penalty function can be approximated by a quadratic function and the whole approximation procedure becomes an iteratively re-weighted least squares problem.

A better approximation suggested by Zou and Li (2008) is the local linear approximation (LLA) algorithm:

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^*|) + \dot{p}_\lambda(|\beta_j^*|)(|\beta_j| - |\beta_j^*|] \qquad for\, \beta_j \approx \beta_j^*. \quad (2.19)$$

With LLA, the penalised likelihood problem become an iteratively re-weighted LASSO problem. According to Fan and Lv (2010), "LLA is a better approximation since it is the minimum (tightest) convex majorant of the concave function on $[0, \infty)$".

3 Feature selection and model specification procedure

3.1 Model description

For any function $f(\cdot)$, throughout this thesis, 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 with 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 as they are known and can be ignored through an affine transformation. 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

$$\ell\big(m(U,X),y\big) \quad \text{with} \quad g\big(m(U,X)\big) = \sum_{j=1}^{d_n} a_j(U)x_j, \tag{3.1}$$

and further assume the support of the index variable U is $[0,1]$ throughout this thesis. Suppose we have a sample (U_i, X_i, y_i), $i = 1, \cdots, n$, from model (3.1), where $X_i = (x_{i1}, \cdots, x_{id_n})^{\mathrm{T}}$. In this chapter, we will introduce how to select the significant variables and identify the constant coefficients in model (3.1), and how to estimate both the functional coefficients and constant coefficients.

It is easy to see, to identify the non-significant variables in (3.1) is equivalent to identify the $a_j(\cdot)$s such that $a_j(U_1) = \cdots = a_j(U_n) = 0$, and to identify the constant coefficients is equivalent to identify the

$a_j(\cdot)$s such that either $\dot{a}_j(U_1) = \cdots = \dot{a}_j(U_n) = 0$ or its deviation $D_j = 0$. The deviation of $a_j(\cdot)$, in this thesis, means the deviation of $a_j(\cdot)$ from its average and is defined as

$$D_j = \left[\sum_{k=1}^{n} \left\{ a_j(U_k) - \frac{1}{n} \sum_{s=1}^{n} a_j(U_s) \right\}^2 \right]^{1/2}. \tag{3.2}$$

Hence, the model selection problem can be transferred to a penalised local maximum likelihood estimation problem. The details of the estimation and model selection procedure are as follows.

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 $\dot{a}_j(U_k)$, $j = 1, \cdots, d_n$,

$$\mathcal{L}_{nk}(\mathbf{a}_k, \mathbf{b}_k) = \frac{1}{n} \sum_{i=1}^{n} \ell \left(g^{-1} \Big\{ \sum_{j=1}^{d_n} \big[\alpha_{jk} + \beta_{jk}(U_i - U_k) \big] x_{ij} \Big\},\ y_i \right) K_h(U_i - U_k), \tag{3.3}$$

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

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

It is easy to see $\mathbf{a}_k$ corresponds to $(a_1(U_k),\ \cdots,\ a_{d_n}(U_k))$ and $\mathbf{b}_k$ corresponds to $(\dot{a}_1(U_k),\ \cdots,\ \dot{a}_{d_n}(U_k))$.

When the number of the covariates is fixed, we may obtain the solutions which maximise the local log-likelihood function $\mathcal{L}_{nk}(\cdot,\cdot)$ defined in (3.3) and show that the resulting nonparametric estimators are consistent (c.f., Cai *et al*, 2000; Zhang and Peng, 2010). However, for the case when the number of covariates is diverging, it would be difficult to obtain satisfactory estimation results by maximising $\mathcal{L}_{nk}(\cdot,\cdot)$ as the number of the unknown nonparametric components involved may exceed the number of observations. In order to address this issue, we next introduce penalised local log-likelihood methods by adding appropriate penalty functions to the above local log-likelihood function. When the number of covariates is diverging, d_n may grow with n in different rate. In this thesis, we call the case diverging dimensional GSVCMs when d_n grows in polynomial rate, and ultra-high dimensional GSVCMs when d_n grows in exponential rate. In Chapter 3.2, we illustrate the feature selection and model specification procedure for diverging dimensional GSVCMs. In Chapter 3.3, we give the feature selection and model specification procedure for ultra-high dimensional GSVCMs.

3.2 Procedure for diverging dimensional GSVCMs

Here we introduce the feature selection and model specification procedure for diverging dimensional GSVCMs. The diverging dimension here means the dimension $d_n \to \infty$ when the sample size $n \to \infty$, and d_n is of order $O(n^{\epsilon_1})$ for some $0 < \epsilon_1 < 1$. The procedure we are going to introduce is a mixture of the ideas of penalised likelihood, local linear approximation and group variable selection. We first use

the local log-likelihood function to construct preliminary estimators of unknown functional coefficients. Then based on the preliminary estimators, we approximate the penalised local log-likelihood function by the sum of a quadratic function and penalties on grouped variables. Therefore we convert the complex penalised local log-likelihood estimation problem to the penalised least-square problem. And with the help of the iterative algorithms provided in Chapter 5.1, the minimisation of the penalised least-square target function can be solved as an iterative re-weighted LASSO problem which we are familiar with.

Let

$$\begin{aligned}
&\mathcal{L}_n(\mathcal{A},\ \mathcal{B}) \\
=&\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n}\ell\left(g^{-1}\Big\{\sum_{j=1}^{d_n}\Big[\alpha_{jk}+\beta_{jk}(U_i-U_k)\Big]x_{ij}\Big\},\ y_i\right)K_h(U_i-U_k) \\
=&\sum_{k=1}^{n}\mathcal{L}_{nk}(\mathbf{a}_k,\mathbf{b}_k), \qquad (3.4)
\end{aligned}$$

where $\mathcal{A}=\big(\mathbf{a}_1^{\mathrm{T}},\ \cdots,\ \mathbf{a}_n^{\mathrm{T}}\big)^{\mathrm{T}}$, $\mathcal{B}=\big(\mathbf{b}_1^{\mathrm{T}},\ \cdots,\ \mathbf{b}_n^{\mathrm{T}}\big)^{\mathrm{T}}$, $\mathbf{a}_k=\big[a_1(U_k),\ \cdots,\ a_{d_n}(U_k)\big]^{\mathrm{T}}$, and $\mathbf{b}_k=\big[\dot{a}_1(U_k),\ \cdots,\ \dot{a}_{d_n}(U_k)\big]^{\mathrm{T}}$ for $k=1,\ \cdots,n$.

The penalised local log-likelihood function for feature selection and structure specification is

$$Q_n(\mathcal{A},\ \mathcal{B})=\mathcal{L}_n(\mathcal{A},\ \mathcal{B})-\sum_{j=1}^{d_n}p_{\lambda_{1j}}\big(\|\boldsymbol{\beta}_j\|\big)-\sum_{j=1}^{d_n}p_{\lambda_{2j}}\big(\|\boldsymbol{\alpha}_j\|\big), \qquad (3.5)$$

where $p_\lambda(\cdot)$ is a penalty function with tuning parameter λ,

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

$\boldsymbol{\alpha}_j$ corresponds to $(a_j(U_1),\ \cdots,\ a_j(U_n))^{\mathrm{T}}$, and $\boldsymbol{\beta}_j$ corresponds to $(\dot{a}_j(U_1),\ \cdots,\ \dot{a}_j(U_n))^{\mathrm{T}}$.

The maximisation of $Q_n(\mathcal{A},\ \mathcal{B})$ can be challenging, and the computation involved could be very expensive. However, by some simple approximations, we could alleviate the computational burden significantly.

Let $(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n)$ be the maximiser of $\mathcal{L}_n(\mathcal{A},\ \mathcal{B})$, and

$$\mathcal{L}_{n*}(\mathcal{A},\mathcal{B}) = \frac{1}{2}\left(\left(\mathcal{A}-\widetilde{\mathcal{A}}_n\right)^{\mathrm{T}},\ h\left(\mathcal{B}-\widetilde{\mathcal{B}}_n\right)^{\mathrm{T}}\right)\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n)\begin{pmatrix}\mathcal{A}-\widetilde{\mathcal{A}}_n\\ h\left(\mathcal{B}-\widetilde{\mathcal{B}}_n\right)\end{pmatrix}.$$

By Taylor's expansion and that $\dot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\widetilde{\mathcal{B}}_n)=0$, we have

$$\mathcal{L}_n(\mathcal{A},\ \mathcal{B}) \approx \mathcal{L}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n) + \mathcal{L}_{n*}(\mathcal{A},\ \mathcal{B}). \tag{3.6}$$

The second derivative $\ddot{\mathcal{L}}_n(\mathcal{A},\ \mathcal{B})$ can be obtained by some tedious computations, it is

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

with

$$\ddot{\mathcal{L}}_n(\mathcal{A},\ \mathcal{B},\ l,\ \kappa) = \mathsf{diag}\left(\ddot{\mathcal{L}}_{n1}(\mathcal{A},\ \mathcal{B},\ l,\ \kappa),\ \cdots,\ \ddot{\mathcal{L}}_{nn}(\mathcal{A},\ \mathcal{B},\ l,\ \kappa)\right),$$

for $l = 0,\ 1,\ \kappa = 0,\ 1$, and

$$\begin{aligned}&\ddot{\mathcal{L}}_{nk}(\mathcal{A},\ \mathcal{B},\ l,\ \kappa)\\ =&\sum_{i=1}^{n} q_2\left(\sum_{j=1}^{d_n}\Big\{\alpha_{jk}+\beta_{jk}(U_i-U_k)\Big\}x_{ij},\ y_i\right)\left(\frac{U_i-U_k}{h}\right)^{l+\kappa} X_i X_i^{\mathrm{T}} K_h(U_i-U_k),\end{aligned}$$

for $k = 1,\ \cdots,\ n$, where $q_2(s,\ y) = \partial^2 \ell\left(g^{-1}(s),\ y\right)/\partial s^2$.

We also apply Taylor's expansion to the penalty functions. By simple calculations, we have

$$p_{\lambda_{1j}}\big(\|\boldsymbol{\beta}_j\|\big) \approx p_{\lambda_{1j}}\left(\|\widetilde{\boldsymbol{\beta}}_j\|\right) - \dot{p}_{\lambda_{1j}}\left(\|\widetilde{\boldsymbol{\beta}}_j\|\right)\|\widetilde{\boldsymbol{\beta}}_j\| + \dot{p}_{\lambda_{1j}}\left(\|\widetilde{\boldsymbol{\beta}}_j\|\right)\|\boldsymbol{\beta}_j\| \tag{3.7}$$

and

$$p_{\lambda_{2j}}\left(\|\boldsymbol{\alpha}_j\|\right) \approx p_{\lambda_{2j}}\left(\|\widetilde{\boldsymbol{\alpha}}_j\|\right) - \dot{p}_{\lambda_{2j}}\left(\|\widetilde{\boldsymbol{\alpha}}_j\|\right)\|\widetilde{\boldsymbol{\alpha}}_j\| + \dot{p}_{\lambda_{2j}}\left(\|\widetilde{\boldsymbol{\alpha}}_j\|\right)\|\boldsymbol{\alpha}_j\|, \tag{3.8}$$

for $j = 1,\ \cdots,\ d_n$.

Let

$$\mathcal{P}_{1n,j}(\|\boldsymbol{\beta}_j\|) = \dot{p}_{\lambda_{1j}}\left(\|\widetilde{\boldsymbol{\beta}}_j\|\right)\|\boldsymbol{\beta}_j\|, \quad \mathcal{P}_{2n,j}(\|\boldsymbol{\alpha}_j\|) = \dot{p}_{\lambda_{2j}}\left(\|\widetilde{\boldsymbol{\alpha}}_j\|\right)\|\boldsymbol{\alpha}_j\|,$$

by (3.6)–(3.8), we define a new objective function

$$Q_{n*}(\mathcal{A},\ \mathcal{B}) = \mathcal{L}_{n*}(\mathcal{A},\ \mathcal{B}) - \sum_{j=1}^{d_n}\mathcal{P}_{1n,j}(\|\boldsymbol{\beta}_j\|) - \sum_{j=1}^{d_n}\mathcal{P}_{2n,j}(\|\boldsymbol{\alpha}_j\|). \tag{3.9}$$

Our feature selection and model specification procedure for diverging dimensional GSVCMs is based on maximising (3.9) rather than (3.5).

Let $(\hat{\boldsymbol{\alpha}}_j,\ \hat{\boldsymbol{\beta}}_j)$, $j = 1,\ \cdots,\ d_n$, be the maximiser of $Q_{n*}(\mathcal{A},\ \mathcal{B})$, and

$$\widehat{\boldsymbol{\alpha}}_j = \big(\widehat{\alpha}_{j1},\ \cdots,\ \widehat{\alpha}_{jn}\big)^{\mathrm{T}}, \quad \widehat{\boldsymbol{\beta}}_j = \big(\widehat{\beta}_{j1},\ \cdots,\ \widehat{\beta}_{jn}\big)^{\mathrm{T}}.$$

$\hat{\boldsymbol{\alpha}}_j$ is our estimator of $(a_j(U_1),\ \cdots,\ a_j(U_n))^{\mathrm{T}}$ and $\hat{\boldsymbol{\beta}}_j$ is our estimator of $(\dot{a}_j(U_1),\ \cdots,\ \dot{a}_j(U_n))^{\mathrm{T}}$, $j = 1,\ \cdots,\ d_n$. Throughout this thesis, we still call $\hat{\boldsymbol{\alpha}}_j$ or $\hat{\boldsymbol{\beta}}_j$ penalised local maximum likelihood estimator, although $Q_{n*}(\mathcal{A},\ \mathcal{B})$ is not the penalised local log-likelihood function.

If we choose an appropriate penalty function, such as SCAD or L_1 penalty, we would expect $\|\widehat{\boldsymbol{\alpha}}_j\| = 0$ when $a_j(\cdot) = 0$ and $\|\widehat{\boldsymbol{\beta}}_j\| = 0$ when $a_j(\cdot)$ is a constant. So, our feature selection and model specification procedure works as follows: if $\|\widehat{\boldsymbol{\alpha}}_j\| = 0$, the corresponding variable x_j is not significant and should be removed from the model. If $\|\widehat{\boldsymbol{\beta}}_j\| = 0$, the coefficient of x_j is constant. Further, when $a_j(\cdot)$ is a constant, denoted by C_j, we use

$$\widehat{C}_j = n^{-1} \sum_{i=1}^{n} \widehat{\alpha}_{ji} \tag{3.10}$$

to estimate C_j.

In the identification of the constant coefficients, an alternative way of penalising the derivatives of the coefficients is to penalise the D_js, defined in (3.2). This leads to the following objective function for feature selection and model specification

$$\tilde{Q}_n(\mathcal{A},\ \mathcal{B}) = \mathcal{L}_{n*}(\mathcal{A},\ \mathcal{B}) - \sum_{j=1}^{d_n} \tilde{\mathcal{P}}_{1n,j}(\mathcal{D}_j) - \sum_{j=1}^{d_n} \mathcal{P}_{2n,j}(\|\boldsymbol{\alpha}_j\|), \tag{3.11}$$

where $\tilde{\mathcal{P}}_{1n,j}(\mathcal{D}_j) = \dot{p}_{\lambda_{1j}}(\widetilde{\mathcal{D}}_j)\mathcal{D}_j$, and $\widetilde{\mathcal{D}}_j$ is $\mathcal{D}_j$ with α_{jk} being replaced by the local maximum likelihood estimator $\tilde{\alpha}_{jk}$ of $a_j(U_k)$, which can be obtained by maximising the local log-likelihood function $\mathcal{L}_n(\mathcal{A},\ \mathcal{B})$ defined in (3.3).

As the maximiser of $\tilde{Q}_n(\mathcal{A},\ \mathcal{B})$ enjoys all asymptotic properties of the maximiser of $Q_{n*}(\mathcal{A},\ \mathcal{B})$, and the theoretical proofs are very similar, in this thesis, we only present the asymptotic properties of the feature selection, model specification and estimation resulting from the maximisation of $Q_{n*}(\mathcal{A},\ \mathcal{B})$.

3.3 Procedure for ultra-high dimensional GSVCMs

We now introduce the feature selection and model specification procedure for ultra-high dimensional GSVCMs. The ultra-high dimension here means the dimension d_n has the exponential order of the sample size n, i.e. $d_n = \exp\{O(n^{\epsilon_2})\}$ for some $\epsilon_2 > 0$. Unlike the diverging dimensional case, as the dimension d_n is allowed to be much larger than the sample size n, we can no longer obtain the preliminary estimators from the local log-likelihood estimation. For the procedure for ultra-high dimensional GSVCMs, we first propose a penalised likelihood method with the LASSO penalty function to get the preliminary estimators of functional coefficients. According to the asymptotic results in Chapter 4.2, the preliminary estimators are uniformly consistent. Then, we use the preliminary estimators of the the functional coefficients to approximate the local log-likelihood function by an L_2 objective function. We also use the preliminary estimators to construct the adaptive group LASSO penalty and the adaptive SCAD penalty.

Hence, we can establish a novel penalised weighted least square method to simultaneously select the significant covariates and identify the constant coefficients among the selected ones. In addition, the algorithm provided in Chapter 5.2 shows this procedure can also be treated as an iterative re-weighted LASSO process and the computation cost is manageable.

Without loss of generality, we assume that there exist $1 \leq s_{n1} \leq s_{n2} < d_n$ such that for $1 \leq j \leq s_{n1}$, $a_j(\cdot)$ are the functional coefficients with non-zero deviation; for $s_{n1}+1 \leq j \leq s_{n2}$, $a_j(\cdot) \equiv c_j$ are the constant coefficients; for $s_{n2}+1 \leq j \leq d_n$, $a_j(\cdot) \equiv 0$. Moreover, we assume that s_{n2}, although may be diverging with the sample size, is much smaller than the sample size n and the number of 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

$$\mathcal{Q}_{nk}(\mathbf{a}_k, \mathbf{b}_k) = \mathcal{L}_{nk}(\mathbf{a}_k, \mathbf{b}_k) - \lambda_1 \sum_{j=1}^{d_n} |\alpha_{jk}| - \lambda_2 \sum_{j=1}^{d_n} h|\beta_{jk}|, \qquad (3.12)$$

where λ_1 and λ_2 are two tuning parameters. We let $(\widetilde{\mathbf{a}}_k,\ \widetilde{\mathbf{b}}_k)$ be the maximiser of $\mathcal{Q}_{nk}(\cdot,\cdot)$, which will be used as the preliminary estimator in the penalised feature selection and model specification procedure

introduced as follows. Let

$$\begin{aligned}
&\mathcal{L}_n(\mathcal{A},\ \mathcal{B}) \\
=&\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n}\ell\left(g^{-1}\Big\{\sum_{j=1}^{d_n}\Big[\alpha_{jk}+\beta_{jk}(U_i-U_k)\Big]x_{ij}\Big\},\ y_i\right)K_h(U_i-U_k) \\
=&\sum_{k=1}^{n}\mathcal{L}_{nk}(\mathbf{a}_k,\mathbf{b}_k), \qquad (3.13)
\end{aligned}$$

where $\mathcal{A} = \big(\mathbf{a}_1^{\mathrm{T}},\ \cdots,\ \mathbf{a}_n^{\mathrm{T}}\big)^{\mathrm{T}}$ and $\mathcal{B} = \big(\mathbf{b}_1^{\mathrm{T}},\ \cdots,\ \mathbf{b}_n^{\mathrm{T}}\big)^{\mathrm{T}}$. In order to conduct the feature selection and model specification for the ultra-high dimensional GSVCMs, we define the following penalised local log-likelihood function:

$$\mathcal{Q}_n(\mathcal{A},\ \mathcal{B}) = \mathcal{L}_n(\mathcal{A},\ \mathcal{B}) - \sum_{j=1}^{d_n} p_{nj}\big(\|\boldsymbol{\alpha}_j\|\big) - \sum_{j=1}^{d_n} p^*_{nj}\big(\|\boldsymbol{\beta}_j\|\big), \qquad (3.14)$$

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

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

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. However, the maximisation of the objective function $\mathcal{Q}_n(\mathcal{A},\ \mathcal{B})$ can be challenging, and the computation involved could be very expensive. Hence we next introduce some simple approximations to the local likelihood function $\mathcal{L}_n(\mathcal{A},\ \mathcal{B})$ and the penalty terms by using the preliminary estimators $\widetilde{\mathbf{a}}_k$ and $\widetilde{\mathbf{b}}_k$ $(k = 1, \cdots, n)$, which could significantly reduce the computational cost.

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

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

$$\ddot{\mathcal{L}}_{nk}(\mathbf{a}_k, \mathbf{b}_k) = \begin{bmatrix} \ddot{\mathcal{L}}_{nk}(\mathbf{a}_k, \mathbf{b}_k, 0) & \ddot{\mathcal{L}}_{nk}(\mathbf{a}_k, \mathbf{b}_k, 1) \\ \ddot{\mathcal{L}}_{nk}(\mathbf{a}_k, \mathbf{b}_k, 1) & \ddot{\mathcal{L}}_{nk}(\mathbf{a}_k, \mathbf{b}_k, 2) \end{bmatrix},$$

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

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 (3.12). Define

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

By Taylor's expansion of the likelihood function defined in (3.13), we

can obtain the following quadratic approximation:

$$
\begin{aligned}
&\mathcal{L}_n(\mathcal{A},\ \mathcal{B}) \\
\approx&\mathcal{L}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n) + \big[\mathcal{V}_n(\mathcal{A}, h\mathcal{B}) - \mathcal{V}_n(\widetilde{\mathcal{A}}_n, h\widetilde{\mathcal{B}}_n)\big]^{\mathrm{T}} \dot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n) + \\
&\frac{1}{2}\big[\mathcal{V}_n(\mathcal{A}, h\mathcal{B}) - \mathcal{V}_n(\widetilde{\mathcal{A}}_n, h\widetilde{\mathcal{B}}_n)\big]^{\mathrm{T}} \ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n)\big[\mathcal{V}_n(\mathcal{A}, h\mathcal{B}) - \mathcal{V}_n(\widetilde{\mathcal{A}}_n, h\widetilde{\mathcal{B}}_n)\big] \\
\equiv&\mathcal{L}_n^{\diamond}(\mathcal{A},\ \mathcal{B}). \qquad (3.15)
\end{aligned}
$$

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 (3.15) than to directly deal with $\mathcal{L}_n(\mathcal{A},\ \mathcal{B})$.

For the penalty functions $p_{nj}(\cdot)$ and $p_{nj}^*(\cdot)$, we consider two possible cases: (i) the adaptive group LASSO penalty, and (ii) the SCAD penalty. Note that identifying the constant coefficients in model (3.1) is equivalent to identifying the $a_j(\cdot)$s such that either $\dot{a}_j(U_1) = \cdots = \dot{a}_j(U_n) = 0$ or its deviation $D_j = 0$, where D_j is defined in (3.2). Using the preliminary estimation results, we can construct the preliminary estimator of D_j:

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

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

For case (i), we define

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

where λ_3 and λ_3^* are two tuning parameters, κ is pre-determined and

can be chosen as 1 or 2 as in the literature, and $\widetilde{\boldsymbol{\alpha}}_j = [\tilde{a}_j(U_1), \cdots, \tilde{a}_j(U_n)]^{\mathrm{T}}$.

For case (ii), we may apply Taylor's expansion to the SCAD penalty function. By a simple calculation on $p_{nj}\left(\|\boldsymbol{\alpha}_j\|\right)$, we have

$$p_{nj}\left(\|\boldsymbol{\alpha}_j\|\right) \approx p_{nj}\left(\|\widetilde{\boldsymbol{\alpha}}_j\|\right) - \dot{p}_{nj}\left(\|\widetilde{\boldsymbol{\alpha}}_j\|\right)\|\widetilde{\boldsymbol{\alpha}}_j\| + \dot{p}_{nj}\left(\|\widetilde{\boldsymbol{\alpha}}_j\|\right)\|\boldsymbol{\alpha}_j\|, \quad (3.16)$$

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

$$\dot{p}_{nj}(z) \equiv \dot{p}_{\lambda_4}(z) = \lambda_4\big[I(z \le \lambda_4) + \frac{(a_0\lambda_4 - z)_+}{(a_0-1)\lambda} I(z > \lambda_4)\big], \quad (3.17)$$

λ_4 is a tuning parameter and $a_0 = 3.7$ as suggested in Fan and Li (2001). For $p^*_{nj}\big(\|\boldsymbol{\beta}_j\|\big)$, we consider the structure:

$$p^*_{nj}\big(\|\boldsymbol{\beta}_j\|\big) = \dot{p}^*_{nj}(|\widetilde{D}_j|)\|h\boldsymbol{\beta}_j\|, \quad (3.18)$$

where $\dot{p}^*_{nj}(\cdot)$ is defined similarly to $\dot{p}_{nj}(\cdot)$ with λ_4 replaced by λ^*_4.

Based on the approximation of $\mathcal{L}_n(\mathcal{A},\ \mathcal{B})$ and the specification of the penalty functions, we may obtain the following two objective functions:

$$\mathcal{Q}^1_n(\mathcal{A},\ \mathcal{B}) = \mathcal{L}^\diamond_n(\mathcal{A},\ \mathcal{B}) - \lambda_3 \sum_{j=1}^{d_n} \|\widetilde{\boldsymbol{\alpha}}_j\|^{-\kappa}\|\boldsymbol{\alpha}_j\| - \lambda^*_3 \sum_{j=1}^{d_n} |\widetilde{D}_j|^{-\kappa}\|h\boldsymbol{\beta}_j\| \quad (3.19)$$

for the adaptive group LASSO penalty; and

$$\mathcal{Q}^2_n(\mathcal{A},\ \mathcal{B}) = \mathcal{L}^\diamond_n(\mathcal{A},\ \mathcal{B}) - \sum_{j=1}^{d_n} \dot{p}_{\lambda_4}(\|\widetilde{\boldsymbol{\alpha}}_j\|)\|\boldsymbol{\alpha}_j\| - \sum_{j=1}^{d_n} \dot{p}_{\lambda^*_4}(|\widetilde{D}_j|)\|h\boldsymbol{\beta}_j\| \quad (3.20)$$

for the SCAD penalty. Note that the two penalty terms in (3.20) are the weighted LASSO penalty functions, where the weights are determined by the derivative of the SCAD penalty using the preliminary estimators $\widetilde{\boldsymbol{\alpha}}_j$ and $\widetilde{D}_j$. Thus, throughout this thesis, we call the penalty functions in (3.20) as the *adaptive SCAD* penalty. The objective functions in (3.19) and (3.20), in some sense, can be seen as the extension of that in Bradic *et al* (2011) from the parametric linear models to the flexible GSVCMs.

Our feature selection and model specification procedure is based on maximising the objective function in either (3.19) or (3.20). Let

$$\widehat{\boldsymbol{\alpha}}_j = \big(\widehat{\alpha}_{j1},\ \cdots,\ \widehat{\alpha}_{jn}\big)^{\mathrm{T}} \ \text{ and } \ \widehat{\boldsymbol{\beta}}_j = \big(\widehat{\beta}_{j1},\ \cdots,\ \widehat{\beta}_{jn}\big)^{\mathrm{T}},\ \ j = 1,\ \cdots,\ d_n, \tag{3.21}$$

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

$$\overline{\boldsymbol{\alpha}}_j = \big(\overline{\alpha}_{j1},\ \cdots,\ \overline{\alpha}_{jn}\big)^{\mathrm{T}} \ \text{ and } \ \overline{\boldsymbol{\beta}}_j = \big(\overline{\beta}_{j1},\ \cdots,\ \overline{\beta}_{jn}\big)^{\mathrm{T}},\ \ j = 1,\ \cdots,\ d_n, \tag{3.22}$$

be the maximisers of $\mathcal{Q}_n^2(\mathcal{A},\ \mathcal{B})$. By choosing the penalty function as the adaptive group LASSO (or 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 feature selection and model 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

$a_j(\cdot)$ is constant which is denoted by c_j and can be estimated by

$$\widehat{c}_j = n^{-1}\sum_{i=1}^{n}\widehat{\alpha}_{ji} \quad \text{or} \quad \overline{c}_j = n^{-1}\sum_{i=1}^{n}\overline{\alpha}_{ji}, \quad j = s_{n1}+1, \cdots, s_{n2}. \tag{3.23}$$

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

4 Asymptotic properties

4.1 Asymptotic properties for diverging dimensional GSVCMs

We are going to present the asymptotic properties of the feature selection, model specification and estimation procedure designed for diverging dimensional GSVCMs. The assumptions and detailed proofs of the following theoretical results can be found in Chapter 9. The assumptions used in this part are mild and justifiable. The detailed discussion of these assumptions can be found in Remark A.1 in Chapter 9.1.

We will start with the uniform consistency of the local maximum likelihood estimators when the number of covariates tends to infinity, followed by the convergence rates of the proposed penalised local maximum likelihood estimators, the sparsity property of the proposed feature selection and structure specification procedure, and the oracle property of the proposed penalised local maximum likelihood estimators. Also we will show the asymptotic normality of the proposed penalised local maximum likelihood estimators.

Let

$$\widetilde{\mathbf{a}}(U_i) = \big(\widetilde{\alpha}_{1i},\ \cdots,\ \widetilde{\alpha}_{d_n i}\big)^{\mathrm{T}} \quad \text{and} \quad \widetilde{\mathbf{b}}(U_i) = \big(\widetilde{\beta}_{1i},\ \cdots,\ \widetilde{\beta}_{d_n i}\big)^{\mathrm{T}}$$

be the local maximum likelihood estimators of

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

respectively. We first present the uniform consistency of $\widetilde{\mathbf{a}}(U_i)$ and $\widetilde{\mathbf{b}}(U_i)$.

Proposition 4.1. Under the Assumptions A1–A5 in Chapter 9.1, we have

$$\sup_{1\le i\le n} \big\|\widetilde{\mathbf{a}}(U_i) - \mathbf{a}(U_i)\big\| = O_P\Big(\sqrt{\frac{d_n \log n}{nh}}\Big) \tag{4.1}$$

and

$$\sup_{1\le i\le n} \big\|\widetilde{\mathbf{b}}(U_i) - \mathbf{b}(U_i)\big\| = O_P\Big(\sqrt{\frac{d_n \log n}{nh^3}}\Big). \tag{4.2}$$

Remark 4.1. *Assumption A3 in Chapter 9.1 guarantees that the maximal distance between two consecutive index variables U_i is only of the order $O_P\big(\frac{\log n}{n}\big)$, see, for example, Janson (1987). Hence, the observed values of U are sufficiently dense on its compact support. In fact, in Chapter 9.2, we prove that the local maximum likelihood estimators are uniformly consistent on the support of U, from which (4.1) and (4.2) can be easily derived. When d_n is fixed, as assumed by Cai* et al *(2000) and Zhang and Peng (2010), the above uniform convergence rate would be reduced to the well-known uniform convergence rate $O_P\big(\sqrt{\frac{\log n}{nh}}\big)$.*

Let $\widehat{\mathbf{a}}(U_i)$ and $\widehat{\mathbf{b}}(U_i)$ be the proposed penalised maximum likelihood estimators of $\mathbf{a}(U_i)$ and $\mathbf{b}(U_i)$. The following proposition gives the convergence rates of $\widehat{\mathbf{a}}(U_i)$ and $\widehat{\mathbf{b}}(U_i)$.

Proposition 4.2. Under the Assumptions A1–A6 in Chapter 9.1, we have

$$\frac{1}{n}\sum_{i=1}^{n} \big\|\widehat{\mathbf{a}}(U_i) - \mathbf{a}(U_i)\big\|^2 = O_P\left(\frac{d_n}{nh}\right) \tag{4.3}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\widehat{\mathbf{b}}(U_i)-\mathbf{b}(U_i)\right\|^2=O_P\left(\frac{d_n}{nh^3}\right). \tag{4.4}$$

Before presenting our main theorems, without loss of generality, we assume that $a_j(\cdot)$ is a function when $j=1,\ \cdots,\ d_n(1)$; $a_j(\cdot)$ is a constant, denoted by C_j, when $j=d_n(1)+1,\ \cdots,\ d_n(2)$; and $a_j(\cdot)=0$ when $j=d_n(2)+1,\ \cdots,\ d_n$, $1\le d_n(1)<d_n(2)<d_n$. Theorem 4.1 below shows the feature selection and model specification procedure based on the maximiser of $Q_{n*}(\mathcal{A},\ \mathcal{B})$ enjoys the property of sparsity.

Theorem 4.1. (Sparsity) *Under the Assumptions A1–A6 in Chapter 9.1, we have*

$$\lim_{n\to\infty}P\left(\max_{d_n(2)+1\le j\le d_n}\|\widehat{\boldsymbol{\alpha}}_j\|=0\right)=1 \tag{4.5}$$

and

$$\lim_{n\to\infty}P\left(\max_{d_n(1)+1\le j\le d_n}\|\widehat{\boldsymbol{\beta}}_j\|=0\right)=1. \tag{4.6}$$

We next investigate the oracle property of the proposed penalised local maximum likelihood estimators.

Let $\widehat{C}_{j,o}$, $j=d_n(1)+1,\ \cdots,\ d_n(2)$, be the estimator of C_j obtained by the standard estimation procedure for the generalised semi-varying coefficient models, see Zhang and Peng (2010), under the assumption that we know $a_j(\cdot)=0$ when $j=d_n(2)+1,\ \cdots,\ d_n$, and $a_j(\cdot)$ is an unknown constant C_j when $j=d_n(1)+1,\ \cdots,\ d_n(2)$.

Let $\widehat{a}_{j,o}(U_i)$, $j=1,\ \cdots,\ d_n(1)$, be the estimator of $a_j(U_i)$ obtained by the standard estimation procedure for generalised varying coefficient models, see Zhang and Peng (2010), under the assumption that we know $a_j(\cdot)=0$ when $j=d_n(2)+1,\ \cdots,\ d_n$, $a_j(\cdot)$ is a constant C_j

when $j = d_n(1) + 1, \ \cdots, \ d_n(2)$, and we also know the true value of C_j. Let

$$\mathbf{D}_n = \Big(\max_{1 \le i \le n} \big|\widehat{\alpha}_{1i} - \widehat{a}_{1,o}(U_i)\big|, \ \cdots, \ \max_{1 \le i \le n} \big|\widehat{\alpha}_{d_n(1)i} - \widehat{a}_{d_n(1),o}(U_i)\big| \Big)^{\mathrm{T}}$$

and

$$\widehat{\mathbf{C}}_o = \Big(\widehat{C}_{d_n(1)+1,o}, \ \cdots, \ \widehat{C}_{d_n(2),o}\Big)^{\mathrm{T}}, \quad \widehat{\mathbf{C}} = \Big(\widehat{C}_{d_n(1)+1}, \ \cdots, \ \widehat{C}_{d_n(2)}\Big)^{\mathrm{T}},$$

where $\widehat{C}_j$ is our estimator of C_j when $a_j(\cdot)$ is an unknown constant C_j, see Chapter 3.2.

Theorem 4.2. (Oracle property) *Under the Assumptions A1–A7 in Chapter 9.1, for any $d_n(1)$-dimensional vector $\mathbf{B}_n$ with $\|\mathbf{B}_n\| = 1$, we have*

$$\sqrt{nh}\mathbf{B}_n^T\mathbf{D}_n = o_P(1), \tag{4.7}$$

and, for any $(d_n(2) - d_n(1))$-dimensional vector $\mathbf{A}_n$ with $\|\mathbf{A}_n\| = 1$, we have

$$n^{1/2}\mathbf{A}_n^T\Big(\widehat{\mathbf{C}} - \widehat{\mathbf{C}}_o\Big) = o_P(1). \tag{4.8}$$

Remark 4.2. *The above theorem indicates that the difference between the proposed penalised local maximum likelihood estimators and those obtained under the oracle assumptions is uniformly asymptotically negligible. Furthermore, as the observed values of U are sufficiently dense on the compact support as discussed in Remark 4.1, such difference is asymptotically negligible uniformly on the support of U. Our theorem*

extends Theorem 2 in Wang and Xia (2009) to semi-parametric setting with the diverging number of the covariates.

Finally, we are going to present the asymptotic normality of the proposed penalised local maximum likelihood estimators by using the oracle property derived in Theorem 4.2.

For any given $u_0 \in [0,1]$, let U_{i_0} be the closest point to u_0. For $j = 1, \cdots, d_n(1)$, we use $\hat{\alpha}_{ji_0}$ defined in Chapter 3.2 to estimate $a_j(u_0)$, and denote it by $\hat{a}_j(u_0)$. Let

$$\widehat{\mathbf{a}}_1(u_0) = \big[\widehat{a}_1(u_0), \cdots, \widehat{a}_{d_n(1)}(u_0)\big]^{\mathrm{T}}, \quad \mathbf{a}_1(u_0) = \big[a_1(u_0), \cdots, a_{d_n(1)}(u_0)\big]^{\mathrm{T}},$$

$$\mathbf{b}_n(u_0) = \frac{1}{2}\mu_2 h^2 \big[\ddot{\mathbf{a}}_1(u_0)\big], \quad \mathbf{\Gamma}_n(u_0) = \mathsf{E}\big[\varrho(U,X)XX^{\mathrm{T}}|U = u_0\big],$$

$$\mathbf{C} = \big[C_{d_n(1)+1},\ \cdots,\ C_{d_n(2)}\big]^{\mathrm{T}},$$

where

$$\mu_k = \int u^k K(u)du, \quad \ddot{\mathbf{a}}_1(u_0) = \big[\ddot{a}_1(u_0),\ \cdots,\ \ddot{a}_{d_n(1)}(u_0)\big]^{\mathrm{T}},$$

$$\varrho(u,x) = -q_2\big[g(m(u,x)), m(u,x)\big].$$

Let

$$X_* = \big(x_1,\ \cdots,\ x_{d_n(1)}\big)^{\mathrm{T}}, \quad X_\diamond = \big(x_{d_n(1)+1},\ \cdots,\ x_{d_n(2)}\big)^{\mathrm{T}},$$

$\mathbf{\Gamma}_{n1}(\cdot)$ and $\mathbf{\Gamma}_{n2}(\cdot)$ be defined as $\mathbf{\Gamma}_n(\cdot)$ with XX^{T} replaced by $X_*X_*{}^{\mathrm{T}}$ and $X_\diamond X_\diamond{}^{\mathrm{T}}$, respectively.

Corollary 4.1. Under the conditions of Theorem 4.2, for any $d_n(1)$-dimensional vector $\mathbf{B}_n$ with $\|\mathbf{B}_n\| = 1$, we have

$$\sqrt{nh}\mathbf{B}_n^{\mathrm{T}}\mathbf{\Gamma}_{n1}^{\frac{1}{2}}(u_0)\Big[\widehat{\mathbf{a}}_1(u_0)-\mathbf{a}_1(u_0)-\mathbf{b}_n(u_0)\Big] \xrightarrow{d} \mathsf{N}\big(0,\ \nu_0 f_U^{-1}(u_0)\big), \quad (4.9)$$

and for any $[d_n(2) - d_n(1)]$-dimensional vector $\mathbf{A}_n$ with $\|\mathbf{A}_n\| = 1$, we have

$$n^{1/2}\mathbf{A}_n^{\mathrm{T}}\mathbf{\Gamma}_n^{-\frac{1}{2}}\Big(\widehat{\mathbf{C}} - \mathbf{C}\Big) \xrightarrow{d} \mathsf{N}(0,\ 1), \quad (4.10)$$

where $f_U(\cdot)$ is the density function of U, and

$$\nu_0 = \int K^2(u)du, \quad \mathbf{\Gamma}_n = \mathsf{E}\big[\mathbf{\Gamma}_{n2}^{-1}(U)\big].$$

Corollary 4.1 shows the proposed penalised local maximum likelihood estimators enjoy asymptotic normality and optimal convergence rate. However, in practice, it would be better to select the model first, then apply the local maximum likelihood estimation to estimate the unknowns in the selected model. This is because, when the sample size is finite, the optimal tuning parameter for model selection is different to that for estimation. It is impossible to pick up the best model and construct the most accurate estimators simultaneously. We have to do the model selection and estimation separately to get the best model and the most accurate estimators. Furthermore, once the model is selected, there is no need use the penalised method. In our simulation studies and real data analysis, we use this two-stage approach to construct estimators, that is to select the model first, then apply the local maximum likelihood estimation to estimate the unknowns in the

selected model.

4.2 Asymptotic properties for ultra-high dimensional GSVCMs

We now present the asymptotic properties of the feature selection, model specification and estimation procedure designed for ultra-high dimensional GSVCMs. The assumptions and detailed proofs of the following theoretical results can be found in Chapter 10. The assumptions used in this part are mild and justifiable. The detailed discussion of these assumptions can be found in Remark B.1 in Chapter 10.1.

Recall the notations $\mathbf{a}_{k0} = \big[a_1(U_k),\ \cdots,\ a_{d_n}(U_k)\big]^{\mathrm{T}}$ and $\mathbf{b}_{k0} = \big[\dot{a}_1(U_k),\ \cdots,\ \dot{a}_{d_n}(U_k)\big]^{\mathrm{T}},\ k = 1, \cdots, n$. We start with the uniform consistency results for their penalised local log-likelihood estimators $\widetilde{\mathbf{a}}_k = \big[\widetilde{a}_1(U_k),\ \cdots,\ \widetilde{a}_{d_n}(U_k)\big]^{\mathrm{T}}$ and $\widetilde{\mathbf{b}}_k = \big[\widetilde{\dot{a}}_1(U_k), \cdots, \widetilde{\dot{a}}_{d_n}(U_k)\big]^{\mathrm{T}}$, which are the maximisers of the objective function (3.12).

Proposition 4.3. Suppose that Assumptions B1–B4 in Chapter 10.1 are satisfied.

(i) If the moment condition (10.1) and Assumption B5 are satisfied with $d_n \propto n^{\tau_1}$, $0 \leq \tau_1 < \infty$, we have

$$\max_{1\leq k\leq n} \|\widetilde{\mathbf{a}}_k - \mathbf{a}_{k0}\| + \max_{1\leq k\leq n} \|h(\widetilde{\mathbf{b}}_k - \mathbf{b}_{k0})\| = O_P(\sqrt{s_{n2}}\lambda_1), \qquad (4.11)$$

where $0 \leq \tau_1 < \infty$ and s_{n2} is the number of the non-zero functional coefficients

(ii) If the moment condition (10.2) and Assumption B5′ are satisfied

with $d_n \propto \exp\{(nh)^{\tau_2}\}$, then (4.11) also holds, where $0 \leq \tau_2 < 1 - \tau_3$ and $0 < \tau_3 < 1$.

The above proposition indicates that the preliminary penalised estimators $\widetilde{\mathbf{a}}_k$ and $\widetilde{\mathbf{b}}_k$ are uniformly consistent, as Assumption B3 in Chapter 10.1 guarantees that the maximal distance between two consecutive index variables U_i is only of 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 (4.11) depends on s_{n2}, the number of the non-zero functional coefficients, and the tuning parameter λ_1. In Assumptions B5 and B5′, we impose some conditions on the relationship between λ_1 and the well-known uniform convergence rate $\left(\frac{\log h^{-1}}{nh}\right)^{1/2}$, and assume that $\lambda_1 \propto \lambda_2$. As a consequence, the influence of $\left(\frac{\log h^{-1}}{nh}\right)^{1/2}$ and λ_2 would be dominated by that of λ_1. It is also interesting to find from the assumptions in Proposition 4.3 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, which is not difficult to understand.

Let $\widehat{\mathcal{A}}_n = \left(\widehat{\mathbf{a}}_1^{\mathrm{T}}, \cdots, \widehat{\mathbf{a}}_n^{\mathrm{T}}\right)^{\mathrm{T}}$ and $\widehat{\mathcal{B}}_n = \left(\widehat{\mathbf{b}}_1^{\mathrm{T}}, \cdots, \widehat{\mathbf{b}}_n^{\mathrm{T}}\right)^{\mathrm{T}}$, where $\widehat{\mathbf{a}}_k = (\widehat{\alpha}_{1k}, \cdots, \widehat{\alpha}_{d_n k})^{\mathrm{T}}$ and $\widehat{\mathbf{b}}_k = (\widehat{\beta}_{1k}, \cdots, \widehat{\beta}_{d_n k})^{\mathrm{T}}$. We define $\mathcal{A}^o = \left[(\mathbf{a}_1^o)^{\mathrm{T}}, \cdots, (\mathbf{a}_n^o)^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\mathcal{B}^o = \left[(\mathbf{b}_1^o)^{\mathrm{T}}, \cdots, (\mathbf{b}_n^o)^{\mathrm{T}}\right]^{\mathrm{T}}$, where the last $(d_n - s_{n2})$ elements of $\mathbf{a}_k^o$ and the last $(d_n - s_{n1})$ elements of $\mathbf{b}_k^o$ are zeros, $k = 1, \cdots, n$, and then denote the biased oracle estimators $\widehat{\mathcal{A}}_n^o = \left[(\widehat{\mathbf{a}}_1^o)^{\mathrm{T}}, \cdots, (\widehat{\mathbf{a}}_n^o)^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\widehat{\mathcal{B}}_n^o = \left[(\widehat{\mathbf{b}}_1^o)^{\mathrm{T}}, \cdots, (\widehat{\mathbf{b}}_n^o)^{\mathrm{T}}\right]^{\mathrm{T}}$, which maximise the objective function $\mathcal{Q}_n^1(\mathcal{A}^o, \mathcal{B}^o)$ when the penalty function is the adaptive group LASSO. Similarly, for $\mathcal{Q}_n^2(\cdot, \cdot)$ when the adaptive SCAD penalty function is used, we let $\overline{\mathcal{A}}_n$ and $\overline{\mathcal{B}}_n$ be the penalised estimated

values, and $\overline{\mathcal{A}}_n^o$ and $\overline{\mathcal{B}}_n^o$ the corresponding biased oracle estimators. The following theorem gives the relation between the penalised estimators which maximise the objective function (3.19) or (3.20) and the corresponding biased oracle estimators.

Theorem 4.3. *Suppose that the conditions in Proposition 4.3 are satisfied.*

(i) When the penalty is chosen as the adaptive group LASSO function and Assumption B6 in Chapter 10.1 is satisfied, with probability approaching one, the maximisers of the objective function $\mathcal{Q}_n^1(\cdot,\cdot)$ defined in (3.19), $(\widehat{\mathcal{A}}_n, \widehat{\mathcal{B}}_n)$, exist and equal to $(\widehat{\mathcal{A}}_n^o, \widehat{\mathcal{B}}_n^o)$. Furthermore,

$$\frac{1}{n}\left\|\widehat{\mathcal{A}}_n^o - \mathcal{A}_0\right\|^2 = \frac{s_{n2}}{nh}, \quad \frac{1}{n}\left\|\widehat{\mathcal{B}}_n^o - \mathcal{B}_0\right\|^2 = \frac{s_{n2}}{nh^3}, \tag{4.12}$$

where $\mathcal{A}_0$ and $\mathcal{B}_0$ are the vectors of the true functional coefficients and their derivative functions, respectively.

(ii) When the penalty is chosen as the adaptive SCAD function and Assumption B6′ in Chapter 10.1 is satisfied, with probability approaching one, the maximisers to the objective function $\mathcal{Q}_n^2(\cdot,\cdot)$ defined in (3.20), $(\overline{\mathcal{A}}_n, \overline{\mathcal{B}}_n)$, exist and equal to $(\overline{\mathcal{A}}_n^o, \overline{\mathcal{B}}_n^o)$. Furthermore, (4.12) still holds when $\widehat{\mathcal{A}}_n^o$ and $\widehat{\mathcal{B}}_n^o$ are replaced by $\overline{\mathcal{A}}_n^o$ and $\overline{\mathcal{B}}_n^o$, respectively.

Theorem 4.3 suggests, using the proposed feature selection and model specification 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 feature selection and model specifica-

tion 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).

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 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 (3.19) and (3.20) are ignored) and the application of (3.23) 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 of the adaptive SCAD penalty function as the case of the adaptive group LASSO penalty function can be derived similarly (with slightly different assumptions). 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 (3.22), 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 (3.23).

Theorem 4.4. *Suppose that the conditions of Theorem 4.3(ii) are*

satisfied. For any s_{n1}*-dimensional vector* $\mathbf{B}_n$ *with* $\|\mathbf{B}_n\| = 1$*, we have*

$$\sqrt{nh}\mathbf{B}_n^T\overline{\mathbf{D}}_n = o_P(1); \tag{4.13}$$

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

$$\sqrt{n}\mathbf{A}_n^T\left(\overline{\mathbf{C}}_n - \mathbf{C}_n^{uo}\right) = o_P(1). \tag{4.14}$$

Theorem 4.4 above 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. 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 $\overline{c}_j$, $j = s_{n1} + 1, \cdots, s_{n2}$.

5 Computational algorithm

5.1 Algorithm for diverging dimensional GSVCMs

As the feature selection and model specification procedure introduced in Chapter 3.2 are based on the maximiser of $Q_{n*}(\mathcal{A},\ \mathcal{B})$ or $\tilde{Q}_n(\mathcal{A},\ \mathcal{B})$. We are going to address how to maximise $Q_{n*}(\mathcal{A},\ \mathcal{B})$ and $\tilde{Q}_n(\mathcal{A},\ \mathcal{B})$.

We first rearrange $\mathcal{L}_{n*}(\mathcal{A},\ \mathcal{B})$ to make it have the standard form for using group LASSO idea. Let

$$\boldsymbol{\theta} = (\boldsymbol{\alpha}_1^{\mathrm{T}},\ \cdots,\ \boldsymbol{\alpha}_{d_n}^{\mathrm{T}},\ \boldsymbol{\beta}_1^{\mathrm{T}},\ \cdots,\ \boldsymbol{\beta}_{d_n}^{\mathrm{T}})^{\mathrm{T}}, \quad \mathbf{T} = (I_n \otimes e_{1,d_n},\ \cdots,\ I_n \otimes e_{d_n,d_n})^{\mathrm{T}},$$

where $e_{k,d}$ is a d-dimensional unit vector with the kth component being 1. It is easy to see

$$\boldsymbol{\theta} = \begin{pmatrix} \mathbf{T}\mathcal{A} \\ \mathbf{T}\mathcal{B} \end{pmatrix}.$$

Let $\tilde{\boldsymbol{\theta}}$ be $\boldsymbol{\theta}$ with $\mathcal{A}$ and $\mathcal{B}$ being respectively replaced by $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}$, and

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

we have

$$\mathcal{L}_{n*}(\mathcal{A},\ \mathcal{B}) = -\frac{1}{2}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta})^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}). \tag{5.1}$$

Maximisation of $Q_{n*}(\mathcal{A},\ \mathcal{B})$

By (5.1), we have

$$-Q_{n*}(\mathcal{A},\ \mathcal{B}) = \frac{1}{2}(\boldsymbol{\eta}-\mathbf{H}\boldsymbol{\theta})^{\mathrm{T}}(\boldsymbol{\eta}-\mathbf{H}\boldsymbol{\theta}) + \sum_{j=1}^{d_n} \tau_{1j}\|\boldsymbol{\beta}_j\| + \sum_{j=1}^{d_n} \tau_{2j}\|\boldsymbol{\alpha}_j\| \triangleq \mathcal{O}(\boldsymbol{\theta}),$$

where

$$\tau_{1j} = \dot{p}_{\lambda_{1j}}\left(\|\widetilde{\boldsymbol{\beta}}_j\|\right), \quad \tau_{2j} = \dot{p}_{\lambda_{2j}}\left(\|\widetilde{\boldsymbol{\alpha}}_j\|\right).$$

So, the maximiser of $Q_{n*}(\mathcal{A},\ \mathcal{B})$ is the minimiser of $\mathcal{O}(\boldsymbol{\theta})$.

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

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

where H_j is the matrix consisting of the $((j-1)n+1)$th to the (jn)th columns of $\mathbf{H}$. That is, for $j = 1,\ \cdots,\ d_n$,

$$\begin{cases} \boldsymbol{\alpha}_j = 0, \quad \text{if } \|H_j^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{-j})\| < \tau_{2j}, \\ \boldsymbol{\alpha}_j = \left(H_j^{\mathrm{T}}H_j + \tau_{2j}\|\boldsymbol{\alpha}_j\|^{-1}I_n\right)^{-1} H_j^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{-j}), \quad \text{otherwise}, \end{cases}$$

and

$$\begin{cases} \boldsymbol{\beta}_j = 0, \quad \text{if } \|H_{j+d_n}^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{-(j+d_n)})\| < \tau_{1j}, \\ \boldsymbol{\beta}_j = \left(H_{j+d_n}^{\mathrm{T}}H_{j+d_n} + \tau_{1j}\|\boldsymbol{\beta}_j\|^{-1}I_n\right)^{-1} H_{j+d_n}^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{-(j+d_n)}), \quad \text{otherwise}, \end{cases}$$

where

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

$\mathbf{0}_n$ is a n-dimensional vector with each component being 0, and

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

This leads to the following iterative algorithm to minimise $\mathcal{O}(\boldsymbol{\theta})$:

(1) Start with $\boldsymbol{\alpha}_j^{(0)} = \tilde{\boldsymbol{\alpha}}_j$ and $\boldsymbol{\beta}_j^{(0)} = \tilde{\boldsymbol{\beta}}_j$, $j = 1,\ \cdots,\ d_n$.

(2) Let the $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$, $j = 1,\ \cdots,\ d_n$, be $\boldsymbol{\alpha}_j^{(k)}$ and $\boldsymbol{\beta}_j^{(k)}$ just after the kth iteration. Update $\boldsymbol{\alpha}_j^{(k)}$ and $\boldsymbol{\beta}_j^{(k)}$ in the $(k+1)$th iteration as follows: for $j = 1,\ \cdots,\ d_n$,

$$\begin{cases} \boldsymbol{\alpha}_j^{(k+1)} = 0, \quad \text{if } \|H_j^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{-j}^{(k)})\| < \tau_{2j}^{(k)}, \\ \boldsymbol{\alpha}_j^{(k+1)} = \left(H_j^{\mathrm{T}} H_j + \tau_{2j}^{(k)} \|\boldsymbol{\alpha}_j^{(k)}\|^{-1} I_n\right)^{-1} H_j^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{-j}^{(k)}), \quad \text{otherwise,} \end{cases}$$

and

$$\begin{cases} \boldsymbol{\beta}_j^{(k+1)} = 0, \quad \text{if } \|H_{j+d_n}^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{-(j+d_n)}^{(k)})\| < \tau_{1j}^{(k)}, \\ \boldsymbol{\beta}_j^{(k+1)} = \left(H_{j+d_n}^{\mathrm{T}} H_{j+d_n} + \tau_{1j}^{(k)} \|\boldsymbol{\beta}_j^{(k)}\|^{-1} I_n\right)^{-1} H_{j+d_n}^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{-(j+d_n)}^{(k)}), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise,} \end{cases}$$

where

$$\tau_{1j}^{(k)} = \dot{p}_{\lambda_{1j}}\left(\|\boldsymbol{\beta}_j^{(k)}\|\right), \quad \tau_{2j}^{(k)} = \dot{p}_{\lambda_{2j}}\left(\|\boldsymbol{\alpha}_j^{(k)}\|\right),$$

$$\boldsymbol{\theta}_{-j}^{(k)} = \Big((\boldsymbol{\alpha}_1^{(k+1)})^{\mathrm{T}},\ \cdots,\ (\boldsymbol{\alpha}_{j-1}^{(k+1)})^{\mathrm{T}},\ \mathbf{0}_n^{\mathrm{T}},\ (\boldsymbol{\alpha}_{j+1}^{(k)})^{\mathrm{T}},\ \cdots,\ (\boldsymbol{\alpha}_{d_n}^{(k)})^{\mathrm{T}},\ (\boldsymbol{\beta}_1^{(k)})^{\mathrm{T}},\ \cdots,\ (\boldsymbol{\beta}_{d_n}^{(k)})^{\mathrm{T}}\Big)^{\mathrm{T}}$$

and

$$\boldsymbol{\theta}^{(k)}_{-(j+d_n)} = \Big((\boldsymbol{\alpha}^{(k+1)}_1)^{\mathrm{T}},\ \cdots,\ (\boldsymbol{\alpha}^{(k+1)}_{d_n})^{\mathrm{T}},\ (\boldsymbol{\beta}^{(k+1)}_1)^{\mathrm{T}},\ \cdots,\ (\boldsymbol{\beta}^{(k+1)}_{j-1})^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}}_n,\ (\boldsymbol{\beta}^{(k)}_{j+1})^{\mathrm{T}},\ \cdots,\ (\boldsymbol{\beta}^{(k)}_{d_n})^{\mathrm{T}} \Big)^{\mathrm{T}}.$$

If

$$\|\boldsymbol{\alpha}^{(k)}_j\| = 0 \quad \text{and} \quad \|H^{\mathrm{T}}_j(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}^{(k)}_{-j})\| > \tau^{(k)}_{2j},$$

we set

$$\boldsymbol{\alpha}^{(k+1)}_j = \left(H^{\mathrm{T}}_j H_j + \tau^{(k)}_{2j} \Delta^{-1} I_n \right)^{-1} H^{\mathrm{T}}_j (\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}^{(k)}_{-j}),$$

where

$$\Delta = \min\{\|\boldsymbol{\alpha}^{(k)}_l\| : \quad \|\boldsymbol{\alpha}^{(k)}_l\| \neq 0,\ l = 1,\ \cdots,\ d_n\}.$$

If

$$\|\boldsymbol{\beta}^{(k)}_j\| = 0 \quad \text{and} \quad \|H^{\mathrm{T}}_{j+d_n}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}^{(k)}_{-(j+d_n)})\| > \tau^{(k)}_{1j},$$

we set

$$\boldsymbol{\beta}^{(k+1)}_j = \left(H^{\mathrm{T}}_{j+d_n} H_{j+d_n} + \tau^{(k)}_{1j} \Delta^{-1}_1 I_n \right)^{-1} H^{\mathrm{T}}_{j+d_n} (\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}^{(k)}_{-(j+d_n)}),$$

where

$$\Delta_1 = \min\{\|\boldsymbol{\beta}^{(k)}_l\| : \quad \|\boldsymbol{\beta}^{(k)}_l\| \neq 0,\ l = 1,\ \cdots,\ d_n\}.$$

(3) If

$$\sum_{j=1}^{d_n} \left\{ \|\boldsymbol{\alpha}_j^{(k)} - \boldsymbol{\alpha}_j^{(k+1)}\| + \|\boldsymbol{\beta}_j^{(k)} - \boldsymbol{\beta}_j^{(k+1)}\| \right\} \tag{5.2}$$

is smaller than a chosen threshold, we stop the iteration, and $(\boldsymbol{\alpha}_j^{(k+1)},\ \boldsymbol{\beta}_j^{(k+1)})$, $j = 1, \cdots, d_n$, is the minimiser of $\mathcal{O}(\boldsymbol{\theta})$. In practice, this threshold is a small enough number (e.g. 10^{-8} in our program). When (5.2) is below this threshold, we believe the estimators after this iteration converge and there is no need to do more iterations.

Maximisation of $\tilde{Q}_n(\mathcal{A},\ \mathcal{B})$

By (5.1), we have

$$-\tilde{Q}_n(\mathcal{A},\ \mathcal{B}) = \frac{1}{2}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta})^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}) + \sum_{j=1}^{d_n} \tilde{\tau}_{1j}\mathcal{D}_j + \sum_{j=1}^{d_n} \tilde{\tau}_{2j}\|\boldsymbol{\alpha}_j\| \stackrel{\triangle}{=} \tilde{\mathcal{O}}(\boldsymbol{\theta})$$

where

$$\mathcal{D}_j^2 = \sum_{k=1}^{n} (\alpha_{jk} - \bar{\alpha}_j)^2 = \boldsymbol{\alpha}_j^{\mathrm{T}} \Xi \boldsymbol{\alpha}_j, \quad \Xi = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}},$$

$\mathbf{1}_n$ is a n-dimensional vector with each component being 1,

$$\tilde{\tau}_{1j} = \dot{p}_{\lambda_{1j}}\left((\widetilde{\boldsymbol{\alpha}}_j^{\mathrm{T}}\Xi\widetilde{\boldsymbol{\alpha}}_j)^{1/2}\right), \quad \tilde{\tau}_{2j} = \dot{p}_{\lambda_{2j}}\left(\|\widetilde{\boldsymbol{\alpha}}_j\|\right).$$

So, the maximiser of $\tilde{Q}_n(\mathcal{A},\ \mathcal{B})$ is the minimiser of $\tilde{\mathcal{O}}(\boldsymbol{\theta})$.

If $\boldsymbol{\theta}$ is a minimiser of $\tilde{\mathcal{O}}(\boldsymbol{\theta})$, then

$$-H_j^{\mathrm{T}}(\boldsymbol{\eta}-\mathbf{H}\boldsymbol{\theta})+\tilde{\tau}_{2j}\|\boldsymbol{\alpha}_j\|^{-1}\boldsymbol{\alpha}_j+\tilde{\tau}_{1j}(\boldsymbol{\alpha}_j^{\mathrm{T}}\Xi\boldsymbol{\alpha}_j)^{-1/2}\Xi\boldsymbol{\alpha}_j=0,\quad\forall\ \Xi\boldsymbol{\alpha}_j\neq 0. \tag{5.3}$$

Hence, if $\boldsymbol{\theta}$ is a minimiser of $\tilde{\mathcal{O}}(\boldsymbol{\theta})$ and $\Xi\boldsymbol{\alpha}_j\neq 0$,

$$\|H_j^{\mathrm{T}}(\boldsymbol{\eta}-\mathbf{H}\boldsymbol{\theta}_{-j})\|=\|H_j^{\mathrm{T}}H_j\boldsymbol{\alpha}_j+\tilde{\tau}_{2j}\|\boldsymbol{\alpha}_j\|^{-1}\boldsymbol{\alpha}_j+\tilde{\tau}_{1j}(\boldsymbol{\alpha}_j^{\mathrm{T}}\Xi\boldsymbol{\alpha}_j)^{-1/2}\Xi\boldsymbol{\alpha}_j\|\geq\tilde{\tau}_{2j}.$$

Therefore, if $\boldsymbol{\theta}$ is a minimiser of $\tilde{\mathcal{O}}(\boldsymbol{\theta})$, there has to be

$$\Xi\boldsymbol{\alpha}_j=0\quad\text{when }\|H_j^{\mathrm{T}}(\boldsymbol{\eta}-\mathbf{H}\boldsymbol{\theta}_{-j})\|<\tilde{\tau}_{2j}.$$

Further, let $\tilde{\mathcal{O}}^{(+j)}(\boldsymbol{\theta})$ be $\tilde{\mathcal{O}}(\boldsymbol{\theta})$ with $\boldsymbol{\alpha}_j$ being replaced by $\alpha\mathbf{1}_n$ and $\mathcal{D}_j$ being replaced by 0, α is a scalar. If $\boldsymbol{\theta}$ is a minimiser of $\tilde{\mathcal{O}}(\boldsymbol{\theta})$ and $\Xi\boldsymbol{\alpha}_j=0$, $\boldsymbol{\theta}$ has to be a minimiser of $\tilde{\mathcal{O}}^{(+j)}(\boldsymbol{\theta})$. So, there has to be

$$-\mathbf{1}_n^{\mathrm{T}}H_j^{\mathrm{T}}(\boldsymbol{\eta}-\mathbf{H}\boldsymbol{\theta})+\tilde{\tau}_{2j}n^{1/2}\mathrm{sign}(\alpha)=0\qquad\forall\ \alpha\neq 0.$$

which leads to that if $\alpha\neq 0$, then

$$\|\mathbf{1}_n^{\mathrm{T}}H_j^{\mathrm{T}}(\boldsymbol{\eta}-\mathbf{H}\boldsymbol{\theta}_{-j})\|=\|\mathbf{1}_n^{\mathrm{T}}H_j^{\mathrm{T}}H_j\mathbf{1}_n\alpha+\tilde{\tau}_{2j}n^{1/2}\mathrm{sign}(\alpha)\|\geq n^{1/2}\tilde{\tau}_{2j}.$$

So, if $\boldsymbol{\theta}$ is a minimiser of $\tilde{\mathcal{O}}(\boldsymbol{\theta})$, we have

$$\boldsymbol{\alpha}_j=0\quad\text{when }\|H_j^{\mathrm{T}}(\boldsymbol{\eta}-\mathbf{H}\boldsymbol{\theta}_{-j})\|<\tilde{\tau}_{2j}.$$

By (5.3), if $\boldsymbol{\theta}$ is a minimiser of $\tilde{\mathcal{O}}(\boldsymbol{\theta})$, when $\Xi\boldsymbol{\alpha}_j\neq 0$, let $\mathbf{P_1}=n^{-1}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$

and

$$\boldsymbol{\theta}_{+j} = (\boldsymbol{\alpha}_1^{\mathrm{T}},\ \cdots,\ \boldsymbol{\alpha}_{j-1}^{\mathrm{T}},\ n^{-1}\boldsymbol{\alpha}_j^{\mathrm{T}}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}},\ \boldsymbol{\alpha}_{j+1}^{\mathrm{T}},\ \cdots,\ \boldsymbol{\alpha}_{d_n}^{\mathrm{T}},\ \boldsymbol{\beta}_1^{\mathrm{T}},\ \cdots,\ \boldsymbol{\beta}_{d_n}^{\mathrm{T}})^{\mathrm{T}},$$

we have

$$\begin{aligned}
&\|H_j^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{+j})\| \\
=&\|H_j^{\mathrm{T}}H_j\Xi\boldsymbol{\alpha}_j + \tilde{\tau}_{2j}\|\boldsymbol{\alpha}_j\|^{-1}\boldsymbol{\alpha}_j + \tilde{\tau}_{1j}(\boldsymbol{\alpha}_j^{\mathrm{T}}\Xi\boldsymbol{\alpha}_j)^{-1/2}\Xi\boldsymbol{\alpha}_j\| \\
=&\|\Xi H_j^{\mathrm{T}}H_j\Xi\boldsymbol{\alpha}_j + \tilde{\tau}_{2j}\|\boldsymbol{\alpha}_j\|^{-1}\Xi\boldsymbol{\alpha}_j + \tilde{\tau}_{1j}(\boldsymbol{\alpha}_j^{\mathrm{T}}\Xi\boldsymbol{\alpha}_j)^{-1/2}\Xi\boldsymbol{\alpha}_j + \\
&\tilde{\tau}_{2j}\|\boldsymbol{\alpha}_j\|^{-1}\mathbf{P}_{\mathbf{1}}\boldsymbol{\alpha}_j + \mathbf{P}_{\mathbf{1}}H_j^{\mathrm{T}}H_j\Xi\boldsymbol{\alpha}_j\| \\
\geq&\|\Xi H_j^{\mathrm{T}}H_j\Xi\boldsymbol{\alpha}_j + \tilde{\tau}_{2j}\|\boldsymbol{\alpha}_j\|^{-1}\Xi\boldsymbol{\alpha}_j + \tilde{\tau}_{1j}(\boldsymbol{\alpha}_j^{\mathrm{T}}\Xi\boldsymbol{\alpha}_j)^{-1/2}\Xi\boldsymbol{\alpha}_j\| \\
\geq&\|\Xi H_j^{\mathrm{T}}H_j\Xi\Xi\boldsymbol{\alpha}_j + \tilde{\tau}_{2j}\|\boldsymbol{\alpha}_j\|^{-1}\Xi\boldsymbol{\alpha}_j + \tilde{\tau}_{1j}(\boldsymbol{\alpha}_j^{\mathrm{T}}\Xi\boldsymbol{\alpha}_j)^{-1/2}\Xi\boldsymbol{\alpha}_j\| \\
\geq&\tilde{\tau}_{1j}.
\end{aligned}$$

A summary of the above argument leads to that if $\boldsymbol{\theta}$ is a minimiser of $\tilde{\mathcal{O}}(\boldsymbol{\theta})$, there has to be

$$\begin{cases}
-H_j^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}) + \tilde{\tau}_{2j}\|\boldsymbol{\alpha}_j\|^{-1}\boldsymbol{\alpha}_j + \tilde{\tau}_{1j}(\boldsymbol{\alpha}_j^{\mathrm{T}}\Xi\boldsymbol{\alpha}_j)^{-1/2}\Xi\boldsymbol{\alpha}_j = 0 \qquad \forall\ \Xi\boldsymbol{\alpha}_j \neq 0, \\
\qquad\qquad \boldsymbol{\alpha}_j = 0, \qquad \text{if } \|H_j^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{-j})\| < \tilde{\tau}_{2j}, \\
\boldsymbol{\alpha}_j = \alpha\mathbf{1}_n, \qquad \text{if } \|H_j^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{+j})\| < \tilde{\tau}_{1j} \text{ and } \|H_j^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{-j})\| \geq \tilde{\tau}_{2j}, \\
\qquad\qquad \boldsymbol{\beta}_j = \left(H_{j+d_n}^{\mathrm{T}}H_{j+d_n}\right)^{-1}H_{j+d_n}^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{-(j+d_n)}).
\end{cases}$$

This leads to the following iterative algorithm to minimise $\tilde{\mathcal{O}}(\boldsymbol{\theta})$:

(1) Start with $\boldsymbol{\alpha}_j^{(0)} = \tilde{\boldsymbol{\alpha}}_j$ and $\boldsymbol{\beta}_j^{(0)} = \tilde{\boldsymbol{\beta}}_j$, $j = 1,\ \cdots,\ d_n$.

(2) Let the $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$, $j = 1,\ \cdots,\ d_n$, be $\boldsymbol{\alpha}_j^{(k)}$ and $\boldsymbol{\beta}_j^{(k)}$ just after the kth iteration. Update $\boldsymbol{\alpha}_j^{(k)}$ and $\boldsymbol{\beta}_j^{(k)}$ in the $(k+1)$th iteration

as follows: for $j = 1, \cdots, d_n$,
if $\|H_j^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{-j}^{(k)})\| < \tilde{\tau}_{2j}^{(k)}$

$$\boldsymbol{\alpha}_j^{(k+1)} = 0;$$

if $\|H_j^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{+j}^{(k)})\| < \tilde{\tau}_{1j}$ and $\|H_j^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{-j}^{(k)})\| \geq \tilde{\tau}_{2j}$,

$$\boldsymbol{\alpha}_j^{(k+1)} = \left(\mathbf{1}_n^{\mathrm{T}} H_j^{\mathrm{T}} H_j \mathbf{1}_n\right)^{-1} \mathbf{1}_n^{\mathrm{T}} H_j^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{-j}^{(k)})\mathbf{1}_n;$$

other situation else,

$$\boldsymbol{\alpha}_j^{(k+1)} = \left(H_j^{\mathrm{T}} H_j + \tilde{\tau}_{2j}^{(k)} \|\boldsymbol{\alpha}_j^{(k)}\|^{-1} I_n + \tilde{\tau}_{1j}^{(k)} \left((\boldsymbol{\alpha}_j^{(k)})^{\mathrm{T}} \Xi \boldsymbol{\alpha}_j^{(k)}\right)^{-1/2} \Xi\right)^{-1} H_j^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{-j}^{(k)}).$$

Under any circumstance

$$\begin{cases} \boldsymbol{\beta}_j^{(k+1)} = 0, \quad \text{if } (\boldsymbol{\alpha}_j^{(k+1)})^{\mathrm{T}} \Xi \boldsymbol{\alpha}_j^{(k+1)} = 0, \\ \boldsymbol{\beta}_j^{(k+1)} = \left(H_{j+d_n}^{\mathrm{T}} H_{j+d_n}\right)^{-1} H_{j+d_n}^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{-(j+d_n)}^{(k)}), \quad \text{otherwise,} \end{cases}$$

where

$$\tilde{\tau}_{1j}^{(k)} = \dot{p}_{\lambda_{1j}}\left(\left((\boldsymbol{\alpha}_j^{(k)})^{\mathrm{T}} \Xi \boldsymbol{\alpha}_j^{(k)}\right)^{1/2}\right), \quad \tilde{\tau}_{2j}^{(k)} = \dot{p}_{\lambda_{2j}}\left(\|\boldsymbol{\alpha}_j^{(k)}\|\right),$$

$$\boldsymbol{\theta}_{+j}^{(k)} = \Big((\boldsymbol{\alpha}_1^{(k+1)})^{\mathrm{T}}, \cdots, (\boldsymbol{\alpha}_{j-1}^{(k+1)})^{\mathrm{T}}, \mathbf{1}_n^{\mathrm{T}}\left(\mathbf{1}_n^{\mathrm{T}} H_j^{\mathrm{T}} H_j \mathbf{1}_n\right)^{-1} \mathbf{1}_n^{\mathrm{T}} H_j^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{-j}^{(k)}),$$
$$(\boldsymbol{\alpha}_{j+1}^{(k)})^{\mathrm{T}}, \cdots, (\boldsymbol{\alpha}_{d_n}^{(k)})^{\mathrm{T}}, (\boldsymbol{\beta}_1^{(k)})^{\mathrm{T}}, \cdots, (\boldsymbol{\beta}_{d_n}^{(k)})^{\mathrm{T}}\Big)^{\mathrm{T}}.$$

If

$$(\boldsymbol{\alpha}_j^{(k)})^{\mathrm{T}} \Xi \boldsymbol{\alpha}_j^{(k)} = 0 \quad \text{and} \quad \|H_j^{\mathrm{T}}(\boldsymbol{\eta} - \mathbf{H}\boldsymbol{\theta}_{+j}^{(k)})\| \geq \tilde{\tau}_{1j},$$

we set

$$\boldsymbol{\alpha}_j^{(k+1)} = \left(H_j^{\mathrm{T}} H_j + \tilde{\tau}_{2j}^{(k)} \tilde{\Delta}^{-1} I_n + \tilde{\tau}_{1j}^{(k)} \tilde{\Delta}_1^{-1/2} \Xi \right)^{-1} H_j^{\mathrm{T}} (\boldsymbol{\eta} - \mathbf{H} \boldsymbol{\theta}_{-j}^{(k)}),$$

where

$$\tilde{\Delta} = \min \left\{ \|\boldsymbol{\alpha}_l^{(k)}\| : \quad \|\boldsymbol{\alpha}_l^{(k)}\| \neq 0,\ l = 1,\ \cdots,\ d_n \right\},$$

and

$$\tilde{\Delta}_1 = \min \left\{ (\boldsymbol{\alpha}_l^{(k)})^{\mathrm{T}} \Xi \boldsymbol{\alpha}_l^{(k)} : \quad (\boldsymbol{\alpha}_l^{(k)})^{\mathrm{T}} \Xi \boldsymbol{\alpha}_l^{(k)} \neq 0,\ l = 1,\ \cdots,\ d_n \right\}.$$

(3) If

$$\sum_{j=1}^{d_n} \left\{ \|\boldsymbol{\alpha}_j^{(k)} - \boldsymbol{\alpha}_j^{(k+1)}\| + \|\boldsymbol{\beta}_j^{(k)} - \boldsymbol{\beta}_j^{(k+1)}\| \right\} \tag{5.4}$$

is smaller than a chosen threshold, we stop the iteration, and $(\boldsymbol{\alpha}_j^{(k+1)},\ \boldsymbol{\beta}_j^{(k+1)})$, $j = 1,\ \cdots,\ d_n$, is the minimiser of $\tilde{\mathcal{O}}(\boldsymbol{\theta})$. In practice, this threshold is a small enough number (e.g. 10^{-8} in our program). When (5.4) is below this threshold, we believe the estimators after this iteration converge and there is no need to do more iterations.

5.2 Algorithm for ultra-high dimensional GSVCMs

First of all, the preliminary estimators $(\widetilde{\mathbf{a}}_k,\ \widetilde{\mathbf{b}}_k)$ are obtained by maximizing $\mathcal{Q}_{nk}(\cdot, \cdot)$ defined in (3.1) for $k = 1, \cdots, n$. By the local quadratic approximation introduced in Chapter 2.2, the penalty functions in (3.1) can be approximated by quadratic functions. Then the

preliminary estimators are obtained by solving an iterative re-weighted least squares problem. The reason we use local quadratic approximation instead of local linear approximation is to reduce the computational burden. The numerical studies show the proposed preliminary estimation works well and the final feature selection, model specification and estimation results are not very sensitive to the choice of the preliminary estimators.

Given the preliminary estimators, the feature selection and model specification procedure proposed in Chapter 3.3 are based on the maximiser of $Q_n^1(\mathcal{A},\ \mathcal{B})$ and $Q_n^2(\mathcal{A},\ \mathcal{B})$. We are going to address how to maximise $Q_n^1(\mathcal{A},\ \mathcal{B})$ and $Q_n^2(\mathcal{A},\ \mathcal{B})$.

We now 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

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

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 kth 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 Chapter 3.3. 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

$$\mathcal{L}_n^*(\mathcal{A},\ \mathcal{B}) = -\frac{1}{2}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta})^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}). \tag{5.5}$$

Maximisation of $Q_n^1(\mathcal{A},\ \mathcal{B})$ and $Q_n^2(\mathcal{A},\ \mathcal{B})$

Given the preliminary 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:

$$\hat{\mathcal{O}}(\boldsymbol{\theta}) \equiv \frac{1}{2}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta})^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}) + \sum_{j=1}^{d_n}\tau_{1j}\|h\boldsymbol{\beta}_j\| + \sum_{j=1}^{d_n}\tau_{2j}\|\boldsymbol{\alpha}_j\|, \tag{5.6}$$

where $\tau_{1j} = \lambda_3^*|\widetilde{D}_j|^{-\kappa}$ and $\tau_{2j} = \lambda_3\|\widetilde{\boldsymbol{\alpha}}_j\|^{-\kappa}$ for $\mathcal{Q}_n^1(\mathcal{A},\ \mathcal{B})$; and $\tau_{1j} = \dot{p}_{\lambda_4^*}(|\widetilde{D}_j|)$ and $\tau_{2j} = \dot{p}_{\lambda_4}(\|\widetilde{\boldsymbol{\alpha}}_j\|)$ 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 $\hat{\mathcal{O}}(\boldsymbol{\theta})$ is

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

where H_j is the matrix consisting of the $((j-1)n+1)$-th to the (jn)-th column of $\mathbf{H}$. Hence, for $j = 1, \cdots, d_n$, we have $\boldsymbol{\alpha}_j = \mathbf{0}_n$ if $\|H_j^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}_{-j})\| < \tau_{2j}$, otherwise

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

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

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

where $\mathbf{0}_n$ is an n-dimensional vector with each component being 0,

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

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

Step 1. Start with $\boldsymbol{\alpha}_j^{(0)} = \tilde{\boldsymbol{\alpha}}_j$ and $\boldsymbol{\beta}_j^{(0)} = \tilde{\boldsymbol{\beta}}_j$, $j = 1, \cdots, d_n$, where $\tilde{\boldsymbol{\alpha}}_j$ and $\tilde{\boldsymbol{\beta}}_j$ are the preliminary estimators of $(a_j(U_1), \cdots, a_j(U_n))^{\mathrm{T}}$ and $(\dot{a}_j(U_1), \cdots, \dot{a}_j(U_n))^{\mathrm{T}}$, respectively.

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 $\|H_j^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}_{-j}^{(k)})\| < \tau_{2j}^{(k)}$, otherwise

$$\boldsymbol{\alpha}_j^{(k+1)} = \left(H_j^{\mathrm{T}} H_j + \tau_{2j}^{(k)}\|\boldsymbol{\alpha}_j^{(k)}\|^{-1} I_n\right)^{-1} H_j^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}_{-j}^{(k)});$$

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

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

where $\tau_{1j}^{(k)}$ is defined as τ_{1j} in (5.6) but with $\widetilde{D}_j$ replaced by $D_j^{(k)}$, $\tau_{2j}^{(k)}$ is defined as τ_{2j} in (5.6) but with $\widetilde{\boldsymbol{\alpha}}_j$ replaced by $\boldsymbol{\alpha}_j^{(k)}$,

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

Furthermore, if $\|\boldsymbol{\alpha}_j^{(k)}\| = 0$ and $\|H_j^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}_{-j}^{(k)})\| > \tau_{2j}^{(k)}$, we set

$$\boldsymbol{\alpha}_j^{(k+1)} = \Big(H_j^{\mathrm{T}} H_j + \tau_{2j}^{(k)} \Delta^{-1} I_n\Big)^{-1} H_j^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}_{-j}^{(k)}),$$

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

$$\boldsymbol{\beta}_j^{(k+1)} = \Big(H_{j+d_n}^{\mathrm{T}} H_{j+d_n} + \tau_{1j}^{(k)} \Delta_1^{-1} I_n\Big)^{-1} H_{j+d_n}^{\mathrm{T}}(\tilde{\boldsymbol{\eta}} - \mathbf{H}\boldsymbol{\theta}_{-(j+d_n)}^{(k)}),$$

with $\Delta_1 = \min\{\|\boldsymbol{\beta}_l^{(k)}\| : \ \|\boldsymbol{\beta}_l^{(k)}\| \neq 0, \ l = 1, \ \cdots, \ d_n\}$.

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 $\big(\boldsymbol{\alpha}_j^{(k+1)}, \ \boldsymbol{\beta}_j^{(k+1)}\big)$,

$j = 1, \cdots, d_n$, is the minimiser of $\hat{\mathcal{O}}(\boldsymbol{\theta})$. In our program, the threshold is chosen as 10^{-8}.

The simulation studies in Chapter 7 will show that the above iterative procedure works reasonably well in the finite sample cases.

6 Selection of tuning parameters

The tuning parameters involved in the proposed feature selection and model specification procedures play a very important role. In this chapter, we will address how to choose these tuning parameters.

6.1 Tuning parameter selection for diverging dimensional GSVCMs

First, in order to reduce the computational cost, we set the bandwidth of the preliminary maximum log-likelihood estimation as $h = 0.6(d_n/n)^{0.2}$. Also the preliminary estimation results are not very sensitive to the choice of the bandwidth.

For the feature selection and model specification procedure based on $Q_{n*}(\mathcal{A},\ \mathcal{B})$ or $\tilde{Q}_n(\mathcal{A},\ \mathcal{B})$, if we use some proper penalty functions that satisfy sparsity, approximate unbiasedness and continuity as introduced in Chapter 2.2, such as SCAD, it would be reasonable to set

$$\lambda_{11} = \lambda_{12} = \cdots = \lambda_{1d_n} = \lambda_1 \quad \text{and} \quad \lambda_{21} = \lambda_{22} = \cdots = \lambda_{2d_n} = \lambda_2.$$

This is because the need of different tuning parameters for different coefficients would be met by the use of a proper penalty function. In fact, in the proposed iterative algorithms in Chapter 5.1, the extent of penalising a coefficient in each iteration is adjusted by its previous value through the derivative of the penalty function. So, from now on, we set $\lambda_{1j} = \lambda_1$ and $\lambda_{2j} = \lambda_2$ for any j, and select λ_1 and λ_2 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 was used. Taking their suggestion, we construct the GIC for model (3.1) as

$$\begin{aligned}\mathrm{GIC}(\lambda_1,\ \lambda_2) = &-2\sum_{i=1}^{n}\ell(\hat{m}(U_i,\ X_i),\ y_i)\\ &+2\ln\{\ln(n)\}\ln(1.028571d_nh^{-1})(k_1+1.028571k_2h^{-1}),\end{aligned} \tag{6.1}$$

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 λ_1 and λ_2. k_1 is the number of significant covariates with constant coefficients obtained based on the tuning parameters λ_1 and λ_2, and k_2 is the number of significant covariates with functional coefficients obtained based on the tuning parameters λ_1 and λ_2. The minimiser of $\mathrm{GIC}(\lambda_1,\ \lambda_2)$ is the selected λ_1 and λ_2.

6.2 Tuning parameter selection for ultra-high dimensional GSVCMs

First, for the preliminary estimation (3.12), the tuning parameters λ_1 and λ_2 are selected through BIC. The bandwidth is set to be $h = 0.75[(\log d_n)/n]^{0.2}$, which satisfies the assumptions in the asymptotic theory. The reason for not using data-driven method to select this bandwidth is to reduce the computational cost. Also the preliminary estimation results are not very sensitive to the choice of the bandwidth.

Then, for the feature selection and model specification procedure based on $Q_n^1(\mathcal{A},\ \mathcal{B})$ or $Q_n^2(\mathcal{A},\ \mathcal{B})$, the tuning parameters λ_3 and λ_3^* or λ_4 and λ_4^* are also selected by the GIC method. Similar to (6.1), we have the following GIC formula

$$\begin{aligned}\mathrm{GIC}(\lambda,\ \lambda^*) = &-2\sum_{i=1}^{n}\ell(\hat{m}(U_i,\ X_i),\ y_i)\\ &+2\ln\{\ln(n)\}\ln(1.028571 d_n h^{-1})(k_1 + 1.028571 k_2 h^{-1}),\end{aligned} \tag{6.2}$$

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 λ_3 and λ_3^* (or λ_4 and λ_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 λ_3 and λ_3^*, while for the maximisation of $\mathcal{Q}_n^2(\mathcal{A},\ \mathcal{B})$, the

minimiser of GIC(λ_4, λ_4^*) is the selected λ_4 and λ_4^*.

7 Numerical studies

We now investigate the finite sample performance of the proposed feature selection, model specification and estimation procedure by some numerical studies. As the procedure developed for the ultra-high dimensional GSVCMs also works well for diverging dimensional case, here we only report the results based on the procedure for ultra-high dimensional GSVCMs. When the dimension d_n is not exceeding the sample size n, two procedures have very close performance while the procedure for diverging dimensional GSVCMs enjoys a faster computational speed. In practice, we suggest to use the procedure for diverging dimensional GSVCMs when $d_n \leq n$ to reduce the computational cost and use the procedure for ultra-high dimensional GSVCMs when $d_n > n$.

Throughout this chapter, we call the procedure based on (3.19) the adaptive group LASSO method and the procedure based on (3.20) the adaptive SCAD method. For the adaptive group LASSO method, the pre-determined parameter κ is chosen to be 1. For the adaptive SCAD method, the SCAD penalty is defined through its derivative as in (3.17). The kernel function used in this chapter is taken to be the Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$. The bandwidth is chosen to be $h = 0.75[(\log d_n)/n]^{0.2}$. The tuning parameters are selected by the data driven approach described in Chapter 6.2.

7.1 Simulation studies

We are going to use three simulated examples to examine the accuracy of the proposed feature selection, model specification and estimation procedure. We will also examine the oracle property of the proposed estimators.

We will start with a simulated example about semi-varying coefficient Poisson regression models, then an example about varying coefficient models and finally an example about varying coefficient Logistic regression models. In Example 7.1, we will examine and compare the proposed adaptive group LASSO method and the adaptive SCAD method about their performance on feature selection, model specification and estimation. We will see the adaptive SCAD method gives slightly better performance under all simulation settings. Thus we will call the adaptive SCAD method "our method" in the following two examples and only compare it with the existing methods. In Example 7.2, we will compare our method with the KLASSO proposed in Wang and Xia (2009) based on varying coefficient models. In Example 7.3, we will compare our method with the methods appear in Lian (2012) based on varying coefficient Logistic regression models. We will see our method outperforms the existing ones.

Example 7.1. 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 $\mathsf{U}[0, 1]$, then generate y_i

based on

$$P(y_i = k) = \frac{\xi_i^k}{k!} e^{-\xi_i}, \quad \log(\xi_i) = \sum_{j=1}^{d_n} a_j(U_i) x_{ij}. \tag{7.1}$$

We set the $a_j(\cdot)$s in (7.1) to be

$$a_1(U) = \sin(2\pi U), \quad a_2(U) = C_2 = 0.6, \quad a_j(U) = 0, \text{ when } j > 2.$$

For sample size $n = 200$ or $n = 300$, and dimension $d_n = 50$, $d_n = 100$, $d_n = 200$, or $d_n = 500$, we apply either the adaptive group LASSO method or the adaptive 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, mean squared error (MSE) of the estimators of the unknown constant coefficients, and the ratios of correct-, under-, over- and other-fitting. The "under-fitting" means that the selected models either miss some significant covariates, or mis-specify some functional coefficients as the constant coefficients. The "over-fitting" means that the selected models either include some insignificant covariates, or mis-specify some constant coefficients as functional. The "other-fitting" means that there exist both under-fitting and over-fitting in the selected models. The "correct" models have to include and only include the true significant variables as well as correctly identify the true structure of the model.

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 adaptive SCAD method work well for feature selection and model specification, and the adaptive SCAD method gives slightly better performance. Table 2 shows that the estimators obtained by either the adaptive group LASSO method or the adaptive SCAD method are doing very well, and their performance is comparable to that of the oracle estimators.

Example 7.2. 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 KLASSO 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 ϵ_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 $2 \leq j_1, j_2 \leq 7$ 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 $\mathsf{B}(4, 1)$, σ is set to be 1.5.

Table 1: The ratios of model selection in 1,000 simulations

Adaptive group LASSO method				
	Correct	Underfitting	Overfitting	Others
n=200, $d_n = 50$	0.971	0.002	0.027	0.000
n=200, $d_n = 100$	0.952	0.007	0.041	0.000
n=200, $d_n = 200$	0.928	0.021	0.049	0.002
n=200, $d_n = 500$	0.892	0.044	0.059	0.005
n=300, $d_n = 50$	0.982	0.001	0.017	0.000
n=300, $d_n = 100$	0.973	0.003	0.024	0.000
n=300, $d_n = 200$	0.946	0.012	0.041	0.001
n=300, $d_n = 500$	0.919	0.019	0.060	0.002
Adaptive SCAD method				
	Correct	Underfitting	Overfitting	Others
n=200, $d_n = 50$	0.979	0.002	0.019	0.000
n=200, $d_n = 100$	0.960	0.006	0.034	0.000
n=200, $d_n = 200$	0.936	0.017	0.046	0.001
n=200, $d_n = 500$	0.902	0.040	0.054	0.004
n=300, $d_n = 50$	0.990	0.001	0.009	0.000
n=300, $d_n = 100$	0.981	0.003	0.016	0.000
n=300, $d_n = 200$	0.956	0.010	0.034	0.000
n=300, $d_n = 500$	0.924	0.017	0.058	0.001

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

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

Table 2: The MISEs and MSEs of the estimators for the functional and constant coefficients

	Adaptive group LASSO		Adaptive SCAD		Oracle Estimators	
			$n = 200$			
	$\widehat{a}_1(\cdot)$	$\widehat{c}_2$	$\overline{a}_1(\cdot)$	$\overline{c}_2$	$a_1^{uo}(\cdot)$	c_2^{uo}
$d_n = 50$	0.0184	0.0045	0.0182	0.0038	0.0170	0.0032
$d_n = 100$	0.0205	0.0076	0.0198	0.0063	0.0170	0.0032
$d_n = 200$	0.0273	0.0133	0.0265	0.0126	0.0170	0.0032
$d_n = 500$	0.0329	0.0175	0.0306	0.0143	0.0170	0.0032
			$n = 300$			
	$\widehat{a}_1(\cdot)$	$\widehat{c}_2$	$\overline{a}_1(\cdot)$	$\overline{c}_2$	$a_1^{uo}(\cdot)$	c_2^{uo}
$d_n = 50$	0.0175	0.0032	0.0172	0.0029	0.0162	0.0027
$d_n = 100$	0.0184	0.0052	0.0181	0.0048	0.0162	0.0027
$d_n = 200$	0.0247	0.0097	0.0232	0.0081	0.0162	0.0027
$d_n = 500$	0.0288	0.0124	0.0274	0.0092	0.0162	0.0027

The MISEs or MSEs of the estimators obtained by either the adaptive group LASSO method or the adaptive SCAD method. $\widehat{a}_1(\cdot)$ *and* $\widehat{c}_2$ *are the estimators obtained by the adaptive group LASSO method,* $\overline{a}_1(\cdot)$ *and* $\overline{c}_2$ *are the estimators obtained by the adaptive SCAD method, and* $a_1^{uo}(\cdot)$ *and* c_2^{uo} *are the unbiased oracle estimators.*

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

$$\text{REE} = 100 \times \frac{\sum_{i=1}^{n}\sum_{j=1}^{d_n}|\hat{a}_j(U_i) - a_j(U_i)|}{\sum_{i=1}^{n}\sum_{i=1}^{d_n}|\hat{a}_{j,o}(U_i) - a_j(U_i)|} \tag{7.2}$$

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. This shows our method is more accurate than the KLASSO on the estimation side. We thus conclude that our method performs better than the KLASSO on both model selection and estimation.

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.

Example 7.3. In this example, we compare the model selection per-

Table 4: Comparison of estimation results between our method and KLASSO

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

formance of our method with the methods proposed in Lian (2012) for generalized 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:

$$E[y_i|X_i] = \frac{\exp\{\sum_{j=1}^{d_n} a_j(U_i)x_{ij}\}}{1+\exp\{\sum_{j=1}^{d_n} a_j(U_i)x_{ij}\}}. \tag{7.3}$$

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 $\mathrm{cov}(x_{ij_1}, x_{ij_2}) = 0.1^{|j_1 - j_2|}$ for any $2 \leq j_1, j_2 \leq d_n$. The index variable U_i, $i = 1, \cdots, n$, are independently generated from the uniform distribution $\mathsf{U}[0, 1]$.

We set the $a_j(\cdot)$s in (7.3) 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 replicates. 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 are correctly selected into the final model while the latter means the average number of insignificant covariates 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.

7.2 Real data analysis

We now apply the adaptive SCAD 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 circulatory and respiratory problems, measurements of pollutants and other environmental factors in Hong Kong. The collected environmental factors are SO_2 (coded by

Table 5: Comparison of model selection between our method and Lian's methods

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 replicates 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).*

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 circulatory and respiratory 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 (7.1), to fit the data. We apply the proposed adaptive SCAD method to identify the significant variables and the nonzero constant coefficients, and estimate the functional or constant coefficients in the selected model.

The selected model is

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

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 significant effects on the number of daily total hospital admissions for circulatory and respiratory problems, and all these variables have time-varying impacts. The estimates of the impacts of these variables are presented in Figure 1.

Figure 1 shows NO_2 always has a positive impact on the daily number of total hospital admissions for circulatory and respiratory problems, and this impact is stronger in winter and spring than 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

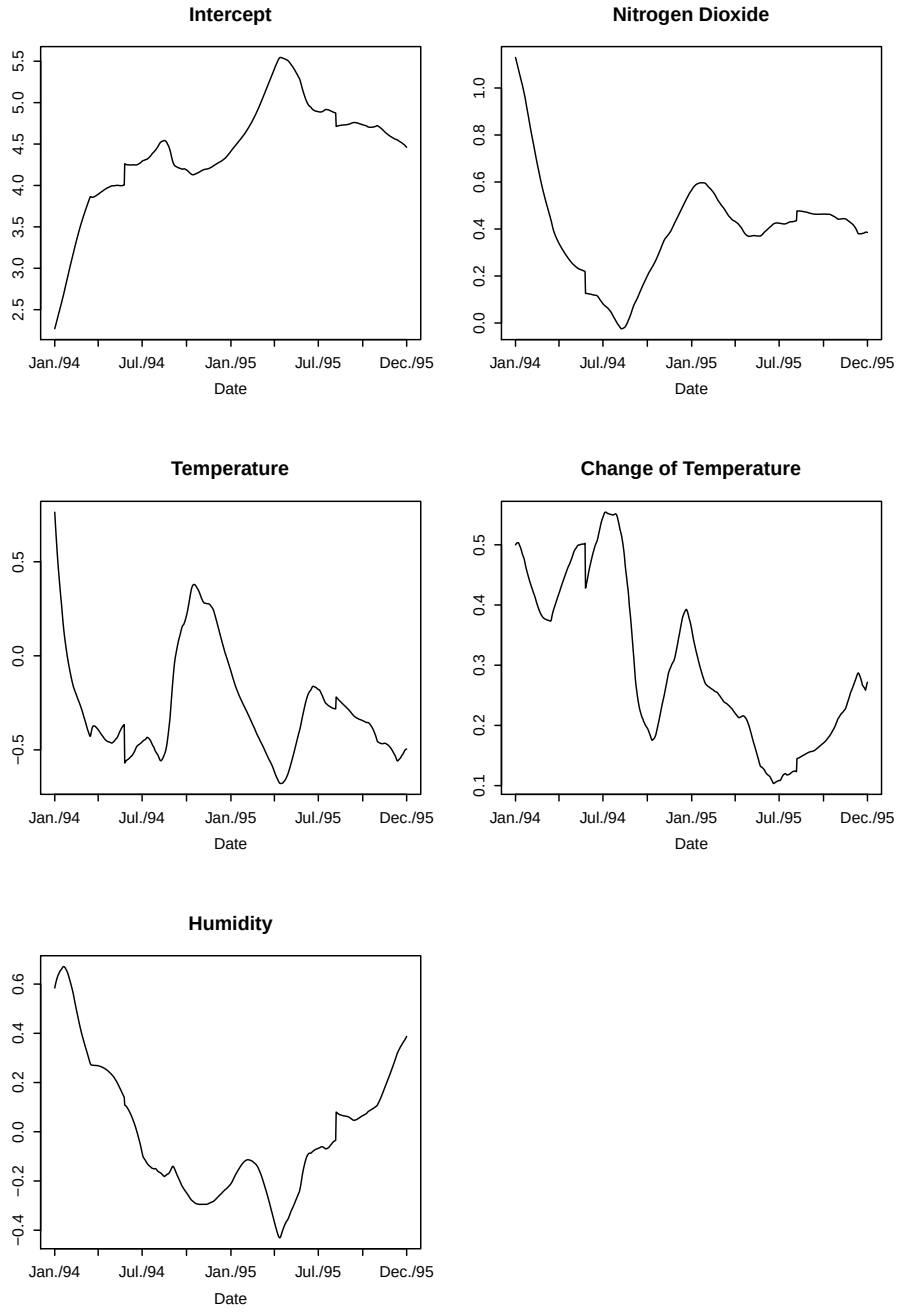
pollution comes from the burning of coals and gasoline. In winter and spring, heating requirement will increase the amount of NO_2 pollution. This is evident from the plot of NO_2 in the data set. Furthermore, the fog and mist in winter and spring will also increase the chance that people expose to NO_2. As a common argument in chemistry (e.g. see Wikipedia), though NO_2 is toxic by inhalation, its compound is acrid and can be easily detected by smell even at low concentrations. Therefore the inhalation exposure to NO_2 can be avoided in most cases. However, when NO_2 is dissolved into the fog, this acid mist will be hard to detect, and people may easily be exposed to this toxic acid mist for a long time without being aware.

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

The impact of humidity on the daily number of total hospital admissions for circulatory and respiratory problems is interesting and complicated. It does not seem to have any seasonal pattern. This is in line with the findings reported in many literature. Indeed, existing researches (Strachan and Sanders, 1989; Schwartz, 1995; and Leon *et al* 1996) agree that humidity has a significant effect on daily hospital

admissions for circulatory and respiratory 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 circulatory and respiratory 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 there may be a strong relationship between humidity level and the risk for children and elderly people to catch circulatory or respiratory disease.

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



8 Multicategory classification method

8.1 Motivation

The motivation for this multicategory classification method arose from a medical study, where the research interest is to classify the patients with early inflammatory polyarthritis at baseline to different risk levels of progression to functional disability in a future time. Such a predictive model is important in stratified medicine in order to identify a sub-group of patients at early stage of disease onset who are at higher risk to progress to a worse outcome so that more aggressive treatment strategies, such as biologic therapy, could be suitable for them. Logistic regression models are widely used for developing predictive models where the outcome of interest is a dischotomous or nominal-scaled variable. When the outcome variable can take more than two values, a multinomial logistic regression is usually applied (Hosmer and Lemeshow, 2000) and new multicategory classification methods in multinomial logistic regression were recently discussed in Li, Jiang, and Fine (2013). However usual logistic regression models assume that the effects of predictors on outcomes are constant. We relax this assumption by allowing the effects of the predictors to vary smoothly with the change of a continuous covariate U based on a varying coefficient structure. The use of such a nonparametric structure permits nonlinear interactions between predictors and a particular variable U and could be useful to improve the model fitting (Fan and Zhang, 1999, 2008; Solari *et al* , 2012). For example, in Chapter 8.4 we will consider a relatively large number of candidate covariates at baseline

to predict future progression to functional disability in early arthritis patients. The set of candidate covariates may include patients' demographic factors (e.g., age, gender), serological and genetic factors (e.g., rheumatic factor status, number of copies of shared epitope), disease activity and severity measures (e.g., number of swollen or tender joints), social-economic factors (e.g., index of multiple deprivation score), etc. The number of potential predictors could be very large if more biomarkers are available at baseline. It is of interest to allow the effects of some baseline predictors to depend on the disease duration from disease symptom onset to the baseline time when predictor variables were measured. By incorporating a flexible interaction between baseline predictors and disease duration in the prognostic model, we account for influences due to variations in time window from disease onset to baseline between subjects. The research questions are then (i) which variables among a large number of candidates should be included in the predictive model; and (ii) which have varying effects among the selected predictors.

In this chapter, we will introduce a multicategory classification method for prognostic classification problems. This method is based on a semi-varying coefficient multinomial logistic regression model and the feature selection and model specification procedure proposed in Chapter 3.2. Similar to the previous chapters, our proposed method allows the number of potential covariates to increase with the sample size and, in theory, tend to infinity as the sample size tends to infinity. This would be particularly useful in practice as we may include all available potential predictors to improve prediction accuracy.

Generally speaking, our proposed multicategory classification method contains three steps:

1. **Feature selection and model specification**. We start with a full model including all potential covariates with functional coefficients and apply a penalised likelihood approach to select predictors and identify which coefficients are functional and which are constant.

2. **Coefficient estimation**. We estimate both constant and functional coefficients based on the selected model.

3. **Classification**. For each subject, we can calculate the conditional probabilities that this subject belongs to different risk groups based on the selected model and its estimated coefficients. The subject is classifiable if the maximum of the estimated group-membership probabilities exceeds a given threshold and is then classified to the corresponding group with the maximum conditional probability. The threshold can be chosen as a high enough probability to distinguish the different groups, for example 80%.

In the above procedures, step 2 is a simple extension of the traditional maximum likelihood estimation and step 3 just involves some trivial calculations. Thus the key step for our modelling is the feature selection and model specification part. The rest of this chapter is arranged as follows. In Chapter 8.2, we discuss the feature selection and model specification using a penalised likelihood estimation method in details, which is based on the procedure we introduced in Chapter 3.2. The bandwidth and tuning parameters for the feature selection and model specification step can be chosen based on the method described in Chapter 6.1. Chapter 8.3 gives simulation studies and Chapter 8.4

focuses on an application to inflammatory polyarthritis data.

8.2 Methodology

A semi-varying coefficient multinomial logistic regression model

Suppose we have a sample $(y_i,\ U_i,\ x_{i1},\ \cdots,\ x_{id_n})$, $i = 1,\ \cdots,\ n$, from $(y,\ U,\ x_1,\ \cdots,\ x_{d_n})$. y is a categorical outcome variable of S levels; U is a given continuous covariate; and x_j, $j = 1,\ \cdots,\ d_n$, are potential predictors that can be either continuous or discrete. We allow d_n to grow and diverge with sample size n. Throughout this chapter, without loss of generality, we assume $y \in \{1,\ \cdots,\ S\}$, and take level S as reference.

Assume the conditional probability that the ith subject belongs to the category s is $p_{si} = P(y_i = s \mid U_i,\ x_{i1}, \ldots, x_{id_n})$, where $i = 1, \ldots, n$ and $s = 1, \ldots, S$. To incorporate nonlinear interactions between x_j and U into the modelling, we specify all p_{si}s through a semi-varying coefficient multinomial logistic regression, i.e.

$$\begin{aligned} p_{si} =& \frac{\exp(\sum_{j=1}^{d_n} x_{ij} a_{sj}(U_i))}{1 + \sum_{k=1}^{S-1} \exp(\sum_{j=1}^{d_n} x_{ij} a_{kj}(U_i))}, \quad s = 1,\ \ldots,\ S-1, \\ p_{Si} =& \frac{1}{1 + \sum_{k=1}^{S-1} \exp(\sum_{j=1}^{d_n} x_{ij} a_{kj}(U_i))}. \end{aligned} \tag{8.1}$$

where $a_{kj}(\cdot)$s are unknown coefficients that are either constant or functional and $\sum_{s=1}^{S} p_{si} = 1$. A constant coefficient $a_{kj}(\cdot)$ means that there is no interaction between x_{ij} and U_i. It follows that the logit of category s versus the reference category S is $\ln\left(\frac{p_{si}}{p_{Si}}\right) = \sum_{j=1}^{d_n} x_{ij} a_{sj}(U_i)$.

Feature selection and model specification

We now describe how to select the predictor variables in (8.1) and identify which coefficients are constant and which are functional. This is basically a feature selection and model specification problem. Based on the penalised likelihood idea, the feature selection and model specification problem is transformed to an estimation problem of the unknown coefficients, $a_{kj}(\cdot)$s, in (8.1). In the following, we are going to apply the penalised local maximum likelihood estimation to estimate $a_{kj}(\cdot)$s in (8.1).

It is easy to see the conditional log-likelihood function of $a_{kj}(\cdot)$s, given all potential predictors, in (8.1) is

$$\sum_{i=1}^{n}\left\{\sum_{s=1}^{S-1} I(y_i = s)\sum_{j=1}^{d_n} x_{ij}a_{sj}(U_i) - \log\left(1+\sum_{k=1}^{S-1}\exp\left\{\sum_{j=1}^{d_n} x_{ij}a_{kj}(U_i)\right\}\right)\right\} \tag{8.2}$$

For each given k, $k = 1, \ldots, n$, within a small neighbourhood of U_k, a Taylor's expansion gives

$$a_{sj}(U_i) \approx a_{sj}(U_k) + \dot{a}_{sj}(U_k)(U_i - U_k),$$

where $i = 1, \ldots, n$, and $j = 1, \ldots, d_n$. This leads to the following local

conditional log-likelihood function

$$\ell_k(\mathbf{a}_k,\ \mathbf{b}_k) = \sum_{i=1}^{n} K_h(U_i - U_k) \left\{ \sum_{s=1}^{S-1} I(y_i = s) \sum_{j=1}^{d_n} x_{ij} \left\{ \alpha_{sjk} + \beta_{sjk}(U_i - U_k) \right\} - \log \left(1 + \sum_{l=1}^{S-1} \exp \left[\sum_{j=1}^{d_n} x_{ij} \left\{ \alpha_{ljk} + \beta_{ljk}(U_i - U_k) \right\} \right] \right) \right\}$$

where α_{sjk} corresponds to $a_{sj}(U_k)$ and β_{sjk} corresponds to $\dot{a}_{sj}(U_k)$, $K(\cdot)$ is a kernel function, h is a bandwidth, $K_h(\cdot) = \frac{1}{h}K(\cdot/h)$,

$$\mathbf{a}_k = (\alpha_{11k},\ \ldots,\ \alpha_{1d_nk},\ \ldots,\ \alpha_{(S-1)1k},\ \ldots,\ \alpha_{(S-1)d_nk})^{\mathrm{T}},$$

$$\mathbf{b}_k = (\beta_{11k},\ \ldots,\ \beta_{1d_nk},\ \ldots,\ \beta_{(S-1)1k},\ \ldots,\ \beta_{(S-1)d_nk})^{\mathrm{T}}.$$

Adding all $\ell_k(\mathbf{a}_k,\ \mathbf{b}_k)$, $k = 1,\ \cdots,\ n$, together, we have

$$\mathcal{L}_n(\mathcal{A},\ \mathcal{B}) = \sum_{k=1}^{n} \ell_k(\mathbf{a}_k,\ \mathbf{b}_k), \tag{8.3}$$

where

$$\mathcal{A} = (\mathbf{a}_1^{\mathrm{T}},\ \ldots,\ \mathbf{a}_{\mathrm{n}}^{\mathrm{T}})^{\mathrm{T}}, \quad \mathcal{B} = (\mathbf{b}_1^{\mathrm{T}},\ \ldots,\ \mathbf{b}_{\mathrm{n}}^{\mathrm{T}})^{\mathrm{T}}.$$

This leads to the following penalised local conditional log-likelihood function for the model selection

$$\mathcal{Q}_n(\mathcal{A},\ \mathcal{B}) = \mathcal{L}_n(\mathcal{A},\ \mathcal{B}) - \sum_{s=1}^{S-1} \sum_{j=1}^{d_n} p_{\lambda_{1sj}}(\mathcal{D}_{sj}) - \sum_{s=1}^{S-1} \sum_{j=1}^{d_n} p_{\lambda_{2sj}}(\|\boldsymbol{\alpha}_{sj}\|), \tag{8.4}$$

where $p_\lambda(\cdot)$ is a penalty function with tuning parameter λ,

$$\|\mathbf{u}\| = (\mathbf{u}^{\mathrm{T}}\mathbf{u})^{1/2}, \quad \boldsymbol{\alpha}_{sj} = (\alpha_{sj1},\ \cdots,\ \alpha_{sjn})^{\mathrm{T}},$$
$$\mathcal{D}_{sj} = \left\{\sum_{k=1}^{n}(\alpha_{sjk} - \bar{\alpha}_{sj})^2\right\}^{1/2}, \quad \text{and } \bar{\alpha}_{sj} = \frac{1}{n}\sum_{k=1}^{n}\alpha_{sjk}.$$

To directly maximise $\mathcal{Q}_n(\mathcal{A},\ \mathcal{B})$ can be very challenging. We are going to find a quadratic function and use its maximiser to approximate the maximiser of $\mathcal{Q}_n(\mathcal{A},\ \mathcal{B})$, thereby simplifying the maximisation.

Let $(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n)$ be the maximiser of $\mathcal{L}_n(\mathcal{A},\ \mathcal{B})$, $\tilde{\alpha}_{sjk}$ the component of $\widetilde{\mathcal{A}}_n$ which corresponds to α_{sjk}. $\widetilde{\boldsymbol{\alpha}}_{sj}$ is $\boldsymbol{\alpha}_{sj}$ with α_{sjk} replaced by $\tilde{\alpha}_{sjk}$. $\widetilde{\mathcal{D}}_{sj}$ is $\mathcal{D}_{sj}$ with α_{sjk} replaced by $\tilde{\alpha}_{sjk}$.

Noticing $\dot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n) = 0$, by Taylor's expansion, we have

$$\begin{aligned}
&\mathcal{L}_n(\mathcal{A},\ \mathcal{B}) \approx \mathcal{L}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n) \\
&+ \frac{1}{2}\left(\left(\mathcal{A} - \widetilde{\mathcal{A}}_n\right)^{\mathrm{T}},\ h\left(\mathcal{B} - \widetilde{\mathcal{B}}_n\right)^{\mathrm{T}}\right)\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n)\begin{pmatrix} \mathcal{A} - \widetilde{\mathcal{A}}_n \\ h\left(\mathcal{B} - \widetilde{\mathcal{B}}_n\right)\end{pmatrix},
\end{aligned}$$

and for $s = 1,\ \cdots,\ S-1$, $j = 1,\ \cdots,\ d_n$,

$$p_{\lambda_{1sj}}(\mathcal{D}_{sj}) \approx p_{\lambda_{1sj}}(\widetilde{\mathcal{D}}_{sj}) - \dot{p}_{\lambda_{1sj}}(\widetilde{\mathcal{D}}_{sj})\widetilde{\mathcal{D}}_{sj} + \dot{p}_{\lambda_{1sj}}(\widetilde{\mathcal{D}}_{sj})\mathcal{D}_{sj},$$

$$p_{\lambda_{2sj}}(\|\boldsymbol{\alpha}_{sj}\|) \approx p_{\lambda_{2sj}}(\|\widetilde{\boldsymbol{\alpha}}_{sj}\|) - \dot{p}_{\lambda_{2sj}}(\|\widetilde{\boldsymbol{\alpha}}_{sj}\|)\,\|\widetilde{\boldsymbol{\alpha}}_{sj}\| + \dot{p}_{\lambda_{2sj}}(\|\widetilde{\boldsymbol{\alpha}}_{sj}\|)\,\|\boldsymbol{\alpha}_{sj}\|.$$

Let

$$\mathcal{L}_{n*}(\mathcal{A},\ \mathcal{B}) = \frac{1}{2}\left((\mathcal{A}-\widetilde{\mathcal{A}}_n)^{\mathrm{T}},\ h(\mathcal{B}-\widetilde{\mathcal{B}}_n)^{\mathrm{T}}\right)\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n)\begin{pmatrix}\mathcal{A}-\widetilde{\mathcal{A}}_n\\ h(\mathcal{B}-\widetilde{\mathcal{B}}_n)\end{pmatrix},$$

and

$$\mathcal{P}_{1n,sj}(\mathcal{D}_{sj}) = \dot{p}_{\lambda_{1sj}}(\widetilde{\mathcal{D}}_{sj})\mathcal{D}_{sj}, \quad \mathcal{P}_{2n,sj}(\|\alpha_j\|) = \dot{p}_{\lambda_{2sj}}(\|\widetilde{\boldsymbol{\alpha}}_{sj}\|)\,\|\boldsymbol{\alpha}_{sj}\|$$

We define

$$\mathcal{Q}_{n*}(\mathcal{A},\ \mathcal{B}) = \mathcal{L}_{n*}(\mathcal{A},\ \mathcal{B}) - \sum_{s=1}^{S-1}\sum_{j=1}^{d_n}\mathcal{P}_{1n,sj}(\mathcal{D}_{sj}) - \sum_{s=1}^{S-1}\sum_{j=1}^{d_n}\mathcal{P}_{2n,sj}(\|\boldsymbol{\alpha}_{sj}\|),$$

and use the maximiser of $\mathcal{Q}_{n*}(\mathcal{A},\ \mathcal{B})$ to approximate the maximiser of $\mathcal{Q}_n(\mathcal{A},\ \mathcal{B})$ and estimate the corresponding unknown parameters.

Let $(\hat{\boldsymbol{\alpha}}_{sj},\ \hat{\boldsymbol{\beta}}_{sj})$, $s = 1,\ \cdots,\ S-1$, $j = 1,\ \cdots,\ d_n$, be the maximiser of $\mathcal{Q}_{n*}(\mathcal{A},\ \mathcal{B})$. For the penalty functions which enjoy sparsity property, such as SCAD or L_1 penalty, our feature selection and model specification procedure works as follows: if $\|\hat{\boldsymbol{\alpha}}_{sj}\| = 0$, then the corresponding variable x_j is not significant and should be removed from modelling the conditional probability $P(y = s|U, x_1, \cdots, x_{d_n})$ of y falling in level s. Let $\hat{\mathcal{D}}_{sj}$ be $\mathcal{D}_{sj}$ with $\boldsymbol{\alpha}_{sj}$ replaced by $\hat{\boldsymbol{\alpha}}_{sj}$. If $\hat{\mathcal{D}}_{sj} = 0$, the coefficient of x_j is constant when modelling $P(y = s|U, x_1, \cdots, x_{d_n})$.

Estimation

After the model is selected, we apply the standard local maximum likelihood estimation to estimate the coefficients based on the selected

model. The details are as following.

Suppose the set of the subscripts of the variables with functional coefficients, in the selected model for $P(y = s|U,\ x_1,\ \cdots,\ x_{d_n})$, is Ω_s, with constant coefficients is Δ_s. For any given u, by simple calculation, we have the following local conditional log likelihood function

$$\sum_{i=1}^{n} K_h(U_i - u) \left\{ \sum_{s=1}^{S-1} I(y_i = s) \left[\sum_{j \in \Omega_s} x_{ij} \left\{ \alpha_{sj} + \beta_{sj}(U_i - u) \right\} + \sum_{l \in \Delta_s} x_{il} \alpha_{sl} \right] \right.$$
$$\left. - \log \left(1 + \sum_{k=1}^{S-1} \exp \left[\sum_{j \in \Omega_k} x_{ij} \left\{ \alpha_{kj} + \beta_{kj}(U_i - u) \right\} + \sum_{l \in \Delta_k} x_{il} \alpha_{kl} \right] \right) \right\}.$$

Let $(\hat{\alpha}_{sj}(u),\ \hat{\beta}_{sj}(u))$, $j \in \Omega_s \cup \Delta_s$, $s = 1,\ \cdots,\ S-1$, be the maximiser of this local conditional log likelihood function at u.

For any $j \in \Omega_s$, the estimator $\hat{a}_{sj}(u)$ of the functional coefficient $a_{sj}(u)$ is taken to be $\hat{\alpha}_{sj}(u)$. For any $l \in \Delta_s$, the coefficient $a_{sl}(\cdot)$ is constant which is denoted by C_{sl}, and can be estimated by

$$\hat{C}_{sl} = \frac{1}{n} \sum_{i=1}^{n} \hat{\alpha}_{sl}(U_i).$$

Classification

Once the model is specified and the coefficients in the selected model are estimated, the classification becomes straightforward: For a new subject, if the observation of the predictor is $(U_l,\ x_{l1},\ \ldots,\ x_{ld_n})$, the conditional probability of this subject falling in level s, $s \in \{1,\ \cdots,\ S-$

1}, given $(U_l,\ x_{l1},\ \ldots,\ x_{ld_n})$ can be estimated by

$$\hat{p}_{sl} = \frac{\exp\left(\sum\limits_{j\in\Omega_s} x_{lj}\hat{a}_{sj}(U_l) + \sum\limits_{j\in\Delta_s} x_{lj}\hat{C}_{sj}\right)}{1 + \sum\limits_{k=1}^{S-1} \exp\left(\sum\limits_{j\in\Omega_k} x_{lj}\hat{a}_{kj}(U_l) + \sum\limits_{j\in\Delta_k} x_{lj}\hat{C}_{kj}\right)}. \tag{8.5}$$

Let

$$\hat{p}_{Sl} = 1 - \sum_{s=1}^{S-1} p_{sl}$$

and $\hat{s}$ maximise $\hat{p}_{sl}$ with respect to s on $\{1,\ \cdots,\ S\}$. If $\hat{p}_{\hat{s}l}$ is greater than a given threshold, this new subject is classifiable under this threshold. Then we classify it into level $\hat{s}$.

8.3 Simulation Study

We are going to use a simulated example to examine the performance of the proposed 3-step multicategory classification method.

Example 8.1. We generate $x_{ij},\ i = 1,\ \cdots,\ n,\ j = 1,\ \cdots,\ d_n$, independently from the standard normal distribution $N(0,\ 1)$, and U_i, $i = 1,\ \cdots,\ n$, from the uniform distribution $U[0,\ 1]$. The response variable y_i is generated from a multinomial logistic regression model defined by (8.1). We set $S = 3$, and all $a_{ij}(\cdot)$s to be 0 except that $a_{11}(U) = \sin(2\pi U)$, $a_{12} = 0.6$, and $a_{21} = 0.7$.

The simulations are conducted for the following cases: the number of potential predictors $d_n = 3, 5, 10, 20$ when the sample size $n = 200$; and $d_n = 50$ when $n = 300$. For each case, we do 200 replicates. In this example, the kernel function is taken to be the

Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$, and the bandwidth is chosen to be $h = 0.6[(S-1)d_n/n]^{0.2}$. We use the SCAD penalty function, which is defined through its derivative as in (3.17). We set the tuning parameters $\lambda_{1sj} = \lambda_1$ and $\lambda_{2sj} = \lambda_2$ all s and j. The tuning parameters λ_1 and λ_2 are selected by the data-driven method introduced in Chapter 6.2.

The performance of the feature selection and model specification step is evaluated by the ratios of picking up the true models at various cases, and the results are presented in Table 6. From Table 6, we can see the feature selection and model specification step works well. The performance of the estimation step is assessed by the median performance among the 200 replicates. The estimated constant coefficients are also reported in Table 6 and the estimates of the functional coefficient $a_{11}(\cdot)$ are presented in Figure 2. From these results, we can see the estimation step works well too.

We then examine the performance of the proposed classification step. We compare the correct classification rate of our method with the oracle rate that based on true model and true coefficients. The correct classification rate is computed by leave-one-out cross-validation. That is: for each i, $i = 1, \cdots, n$, we classify the ith subject by apply the proposed method to the rest $n - 1$ subjects. Then, for each replicate, the correct classification rate is the ratio between the number of correct classified subjects and the number of total subjects (i.e. n). By repeating this calculation for 200 replicates, we can get a sample of correct classification rate. Similarly, we can also get a synthetic sample of oracle rate by classify each subject according to the true

conditional probabilities of all levels. The sample means and sample variances of the correct classification rate and the oracle rate under all cases are presented in Table 7. From Table 7 we can see the proposed classification method performs almost equally well as the oracle one. It should be noted that the proposed classification method does not have a high correct classification rate. This is because, for each case, none of the levels, which the outcome variable y may take, stands out with high conditional probability. This means even if we knew the true conditional probability (like the oracle one), we would still have a good chance to mis-classify a subject. Indeed, from Table 7 one can see, the oracle one does not have high correct classification rate either.

Furthermore, we will demonstrate that the proposed classification method would have high correct classification rate when there is one level standing out with a high conditional probability. Under the same simulation settings, for each case, we treat the simulated sample as a training set, then simulate a test observation. The test observation is simulated such that there is one level which the outcome variable y falls into with conditional probability of either 90% to 100%, or 80% to 90%, or 70% to 80%, or 60% to 70%. We apply the proposed classification method based on the training set to classify the test observation. For each case, we do 200 replicates, and compute the correct classification rate across the 200 replicates. The results are reported in the right hand side of Table 7.

Table 6: Simulation study – ratios of picking up models and estimates of constant coefficients

	Ratios of picking up models				Coefficient estimates	
Cases	Correct	Under	Over	Others	$\hat{C}_{12}$	$\hat{C}_{21}$
n = 200						
$d_n = 3$	0.955	0.015	0.030	0.000	0.540	0.757
$d_n = 5$	0.940	0.025	0.035	0.000	0.678	0.795
$d_n = 10$	0.920	0.025	0.055	0.000	0.513	0.621
$d_n = 20$	0.895	0.035	0.070	0.000	0.723	0.817
n = 300						
$d_n = 50$	0.820	0.065	0.110	0.005	0.466	0.573

Columns corresponding to "Correct", "Under", "Over" and "Others" are the ratios of picking up correct, under-fitting, over-fitting and other models, respectively. True values of constant coefficients are $C_{12} = 0.6$ *and* $C_{21} = 0.7$.

Table 7: Simulation study – comparison of means and variances of correct classification rates between oracle method and our method

	Oracle method		Our method					
Cases	Mean	Variance	Mean	Variance	90%− 100%	80%− 90%	70%− 80%	60%− 70%
n = 200								
$d_n = 3$	0.452	0.001	0.447	0.001	0.920	0.815	0.725	0.610
$d_n = 5$	0.450	0.001	0.436	0.001	0.910	0.805	0.720	0.605
$d_n = 10$	0.448	0.001	0.425	0.002	0.900	0.795	0.710	0.595
$d_n = 20$	0.441	0.001	0.406	0.002	0.890	0.780	0.695	0.585
n = 300								
$d_n = 50$	0.458	0.001	0.368	0.002	0.850	0.740	0.635	0.530

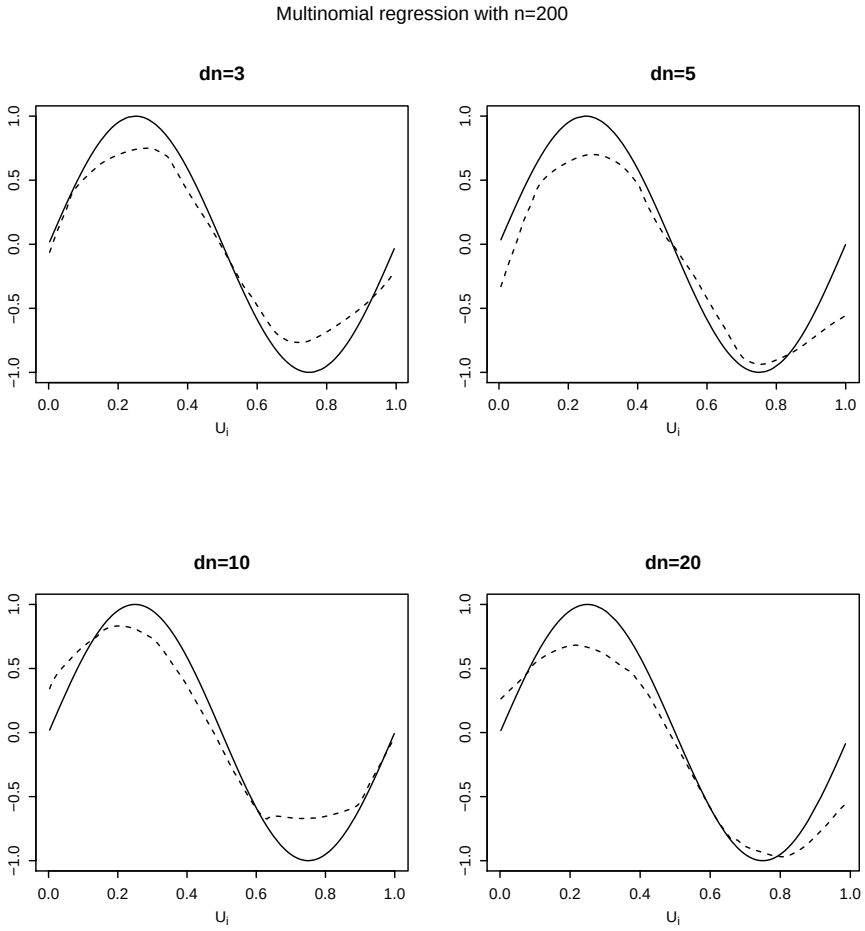
Columns corresponding to "90%-100%", "80%-90%", "70%-80%", "60%-70%" are the average correct classification rates for the sub-group of subjects with maximum conditional probability between 90% *to* 100% *, "80%-90%", "70%-80%", "60%-70%".*

8.4 Application to a medical data set about inflammatory polyarthritis

Scientific background

We now use the classification method developed in Chapter 8.2 to study a medical data set from the primary care-based prospective cohort of

Figure 2: Estimation of functional coefficient in simulation study – solid lines are the true functional coefficients and dotted lines are their estimates.



patients with recent onset inflammatory polyarthritis (Farragher *et al*, 2010). Rheumatoid arthritis (RA) is the most common inflammatory disease of the joints, which is associated with progressive joint destruction resulting in severe disability. However, it is difficult to identify RA at an early stage of disease onset because no tests or di-

agnostic criteria are available to define early RA (Visser, 2005). A lab test that often helps with diagnosis of RA at a follow up stage is anti-cyclic citrullinated peptide antibody test. Early arthritis may be progressed into established RA or another definite arthritis disease or may remain undifferentiated. To better manage the outcome in arthritis, it has been suggested by clinical researchers to first recognize inflammatory arthritis and then estimate the risk of developing persistent and erosive irreversible arthritis such as RA in order to propose an optimal treatment (Dixon and Symmons, 2005). RA is a very heterogeneous disease in terms of disease progression outcome. Some RA patients do not develop any severe outcome, such as erosion, even after a long time, but the majority will have bone erosions and cartilage breakdown resulting in joint destruction and functional disability. For the management of early inflammatory polyarthritis, the European League against Rheumatism recommends that patients at risk of developing persistent and/or erosive arthritis should be started with disease-modifying anti-rheumatic drugs (DMARDs) as early as possible, even if they do not yet fulfil established classification criteria for RA (Combe *et al*, 2007). Furthermore, the revolutionary introduction of biologic agents, such as anti-tumour necrosis factor (TNF), in the past decade offers patients a new and very effective treatment option alternative to the traditional DMARDs. Early treatment with biologic agents has been shown by published studies to improve clinical outcomes, patients' functional status and health-related quality of life (Venkateshan *et al*, 2009). However, biologic agents have potential to leave the patients more vulnerable to severe adverse events such as

infection or malignancy because TNF is involved in many aspects of host immunity (Fu *et al*, 2013) . Also the drug costs of treatment with biologic agents are much higher comparing to DMARDs. In order to achieve the goal of personalised treatment and optimal early use of biologic agents in the management of RA, it is necessary to identify a sub group of patients at baseline who are at higher risk to progress into a worse functional status in future or have better response to biologic treatment so that specific treatment strategies are matched to individual patients. In this study, our scientific interest focuses on a prognostic model to classify the patients into groups with different risk of progression to severe outcome rather than different responses to treatment. An ideal therapeutic strategy should then be based on such an appropriate prognostication of the disease (Combe *et al*, 2007). The aim of this study is to improve the prognostic (or predictive) modelling by identifying significant prognostic factors (or predictors) associated with disease progression together with their significant interactions.

HAQ progression data

The data sample we study comprises 290 patients, who were recruited to the Norfolk arthritis register cohort between 1990-1994 and have disease duration from symptom onset to registration less than three years. The disease outcome of interest is functional disability status, which is an important clinical measure in RA as it has been shown to be predictive of crucial RA-related outcomes, such as mortality (Fang *et al*, 2014). This measure was assessed using the modified British version of the Health Assessment Questionnaire (HAQ) score. The

questionnaire contains 20 questions in 8 categories. Each question is given a score of 0 (no difficulty), 1 (some difficulty), 2 (much difficulty or need of assistance), or 3 (unable to perform). The score for each category is determined by the highest score in that category, and the sum of scores is then divided by the number of categories, yielding a total HAQ score ranging from 0 (best) to 3 (worst). All patients in our study sample have mild disease outcome at registration (baseline) with baseline HAQ scores between 0 and 1 and were followed for at least five years. The response variable Y is the functional disability status at the end of a 5-year follow up since registration. $Y = 1$ if the functional disability status at the end of follow up is at low risk (HAQ score between 0 and 1); 2 if the functional disability status is at moderate risk (HAQ score between 1 and 2); and 3 if the functional disability status is at high risk (HAQ score between 2 and 3).

In the predictive model, the candidate predictors include age at registration, gender, number of swollen joints out of 51 joints, number of tender joints of 51 joints, rheumatic factor (1=positive or 0= negative), smoking status (three categories: non-smoker, current smoker or ex-smoker), socio-economic status defined as a area-level category variable based on the nationally-determined quartiles of the index of multiple deprivation score used in the UK (four categories: least deprived group, two middle deprived groups, most deprived group), number of copies of the shared epitope which is an established genetic biomarker in RA, fulfillment of the American College of Rheumatology 1987 classification criteria for rheumatoid arthritis (1= yes or 0 = no), season of birth (four categories: spring, summer, autumn or winter), DMARDs

treatment duration (in days), baseline HAQ score and their functional interactions with disease duration from symptom onset (in months).

One of the advantages of the proposed semi-varying coefficient multinomial logistic regression model is to allow us to incorporate potentially varying effects of baseline covariates with the change of disease duration on disease progression outcome. It is more flexible and more general than the one including a linear interaction term between a covariate and disease duration. We use the proposed method in Chapter 8.2 to do feature selection and model specification to decide which covariates are those with a varying coefficient and then estimate both constant coefficients and varying coefficients based on the selected model.

Results and analysis

We fit model (8.1) to the data and consider the disease duration variable as the covariate U. Without loss of generality, we re-scale the covariate U to $[0, 1]$. The response category $Y = 1$ is chosen as the reference and the other two categories are compared against the reference category $Y = 1$. The initial model contains 18 covariates with varying coefficients including an intercept and all numerical or dummy variables listed in the data description above. We use the classification method introduced in Chapter 8.2 to do feature selection and model specification. The kernel function is chosen as the Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$, and the bandwidth is chosen as $h = 0.6[(S-1)d_n/n]^{0.2}$. The tuning parameters λ_1 and λ_2 are selected by GIC (Fan and Tang, 2013).

The selected predictors together with their estimated constant or functional coefficients and associated standard errors are presented in Table 8 and Table 9. The plots of the estimation results of functional coefficients are presented in Figure 3 and 4. For those functional coefficients, we report their estimates and standard errors at a given number of U values with disease duration being 1, 3, 6, 12 or 24 months. The standard errors of the coefficient estimates are calculated by a bootstrap method as follows. For each observation, using the estimated coefficients of the selected predictors, we can generate a bootstrap sample member. Then all bootstrap sample members, i.e. bootstrap sample members for all observations, form a bootstrap sample. Based on the bootstrap sample, we can get the estimates of the coefficients, which we call a bootstrap sample member of the estimated coefficients. Repeating the re-sampling procedure 500 times, a bootstrap sample of the estimated coefficients of size 500 is obtained. The sample standard deviation of the bootstrap sample of estimated coefficients is used as the standard error of the estimate.

Among the list of candidate covariates, twelve were selected to be significantly associated with the multinomial logit of the response group $Y = 2$ (moderate risk) relative to the reference group $Y = 1$ (low risk). Three of them (RA, female, current smoker) have constant coefficients and are associated with increasing probability of being a higher risk group. The others (baseline HAQ score, number of swollen joints, number of tender joints, DMARDs treatment duration, age at onset, copies of genetic biomarker, previous smoker, upper middle deprived group, and most deprived group) have functional coefficients. For the

multinomial logit of the response group $Y = 3$ (high risk) relative to the low risk group, eight covariates together with a functional intercept were selected in the model. Two of them (baseline HAQ, rheumatic factor) have constant coefficients and six (number of swollen joints, number of tender joints, DMARDs treatment duration, age at onset, copies of genetic biomarker, upper middle deprived group) have functional coefficients. All the selected covariates in Table 8 and Table 9 are indeed well acknowledged predictors in HAQ progression (see for example, Combe *et al*, 2003). Less significant covariates were identified in Table 9 due to the smaller sample size of 23 in Group $Y = 3$ comparing to 74 in Group $Y = 2$.

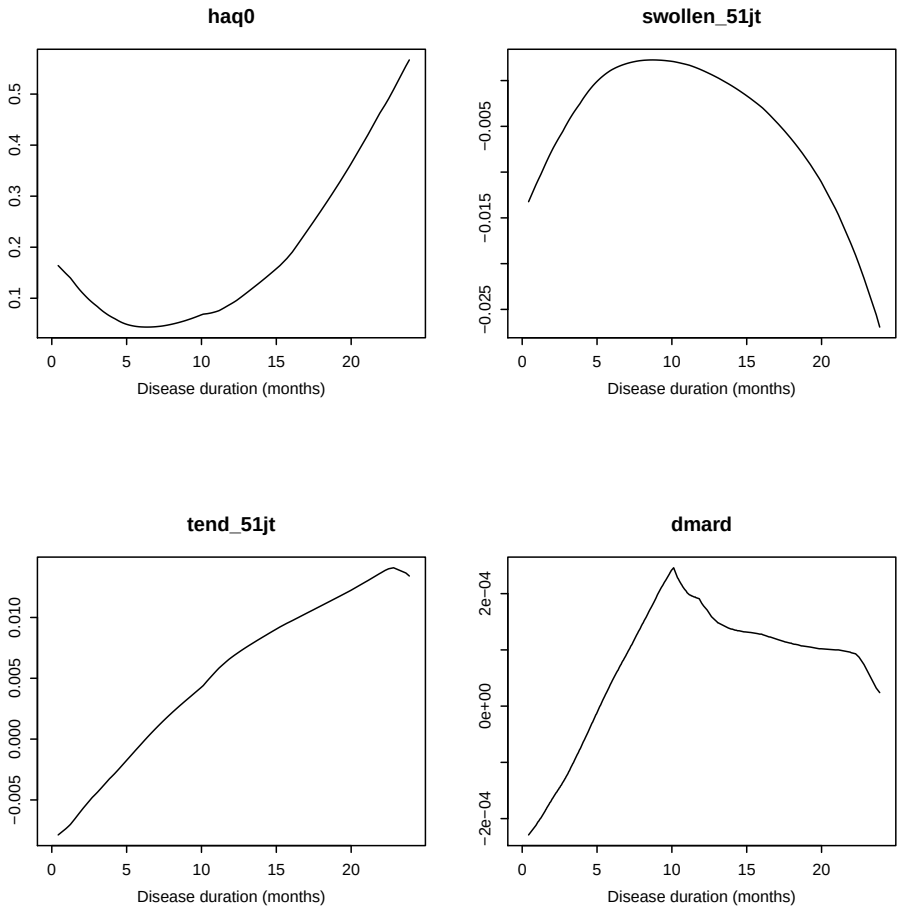
The scientific aim of this study is to classify the patients into different risk groups at baseline to predict their outcomes at the end of follow up. Hence we assess the performance of our methods by comparing correct classification rates with other existing methods. The calculation of correct classification rate is based on a leave-one-out cross-validation approach. For each subject, we use the rest of the data (289 subjects) to select covariates and obtain their coefficient estimates. Then we calculate the estimated conditional probability of belonging to each risk level for this subject. If any estimated conditional group-membership probability is higher than a threshold, say 80% or 70%, we classify this subject into the corresponding group and compare the classification result with the true response value of Y. By repeating this procedure to all subjects, we calculate correct classification rates for those subjects who have a maximum of the estimated group-membership probabilities greater than the threshold. The re-

sults are shown in Table 10, where the correct classification rate are compared between the model we selected (Model 1) and alternatives that can be handled by existing R packages(Models 2-5). We pick Models 2-5 to represent the model structure that are commonly used in the applied fields as they can be solved without extra efforts on programming. Model 2 is the full model including all covariates with functional coefficients; Model 3 is the one including those covariates identified in Table 8 but with constant coefficients; Model 4 is the one including those covariates identified in Table 9 but with constant coefficients; Model 5 is the full model including all covariates with constant coefficients. We see that our selected model always gives the highest correct classification rate comparing to the others. It could reach to 85.2% when the threshold probability is 0.8, though unsurprisingly less number of subjects are classifiable. The two full models, with either constant coefficient or functional coefficient, give lower classification rates due to overfitting.

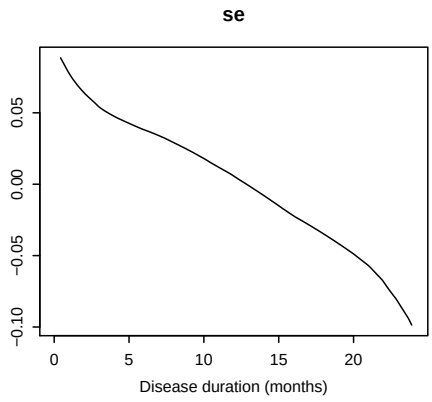
8.5 Discussion

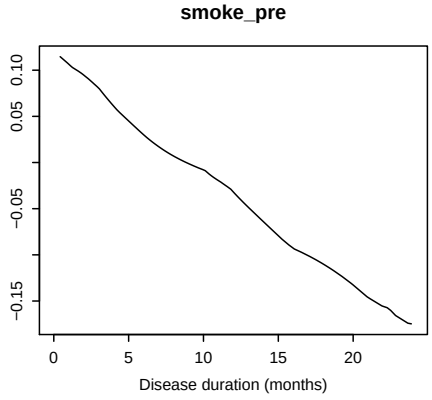
In stratified medicine, it is of interest to identify a sub-group of subjects at baseline who are at high risk in future progression to a severe disease outcome and hence specific therapeutic strategy could be matched. Many prognostic markers (predictors) are often taken into account in prognostic classification modelling and the interactions between predictor variables can be complicated. In this chapter, we presented a semi-varying coefficient regression model for improving the classification in predictive modelling and conducted the feature

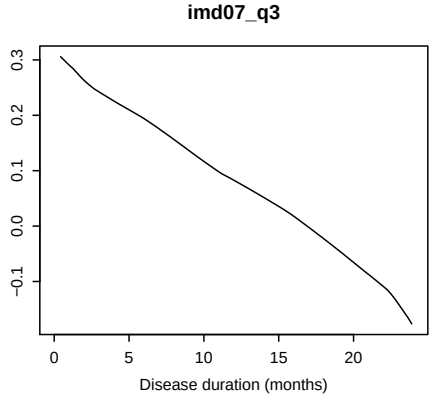
Figure 3: Plots of estimates of functional coefficient versus disease durations – HAQ progression data with selected covariates for the logit of $Y = 2$ vs $Y = 1$.




age
0.0000
−0.0010
−0.0020
0
5
10
15
20
Disease duration (months)


se
0.05
0.00
−0.05
−0.10
0
5
10
15
20
Disease duration (months)


smoke_pre
0.10
0.05
−0.05
−0.15
0
5
10
15
20
Disease duration (months)


imd07_q3
0.3
0.2
0.1
0.0
−0.1
0
5
10
15
20
Disease duration (months)

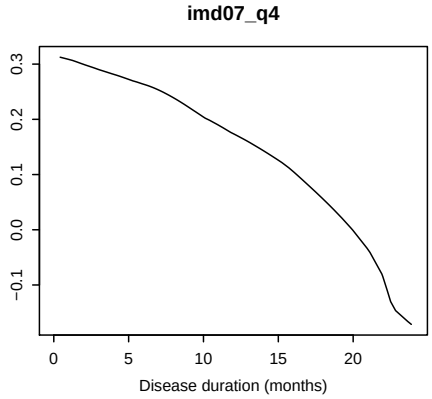
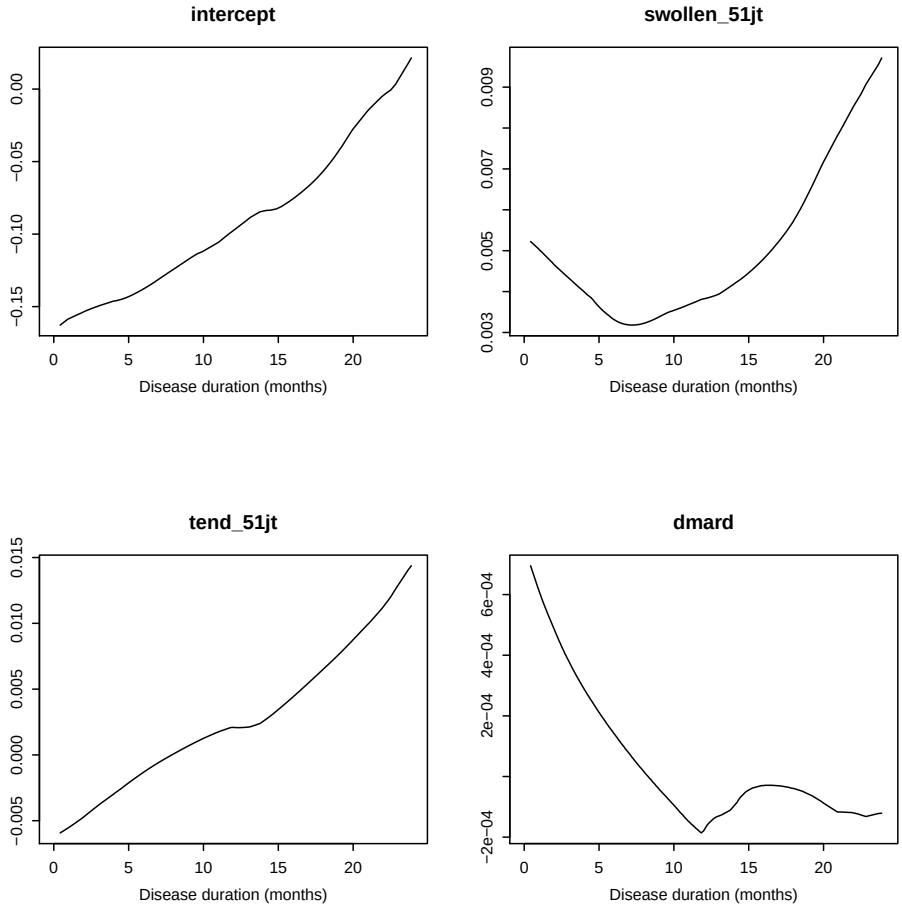
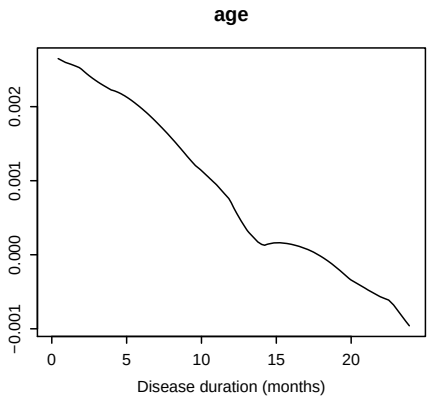
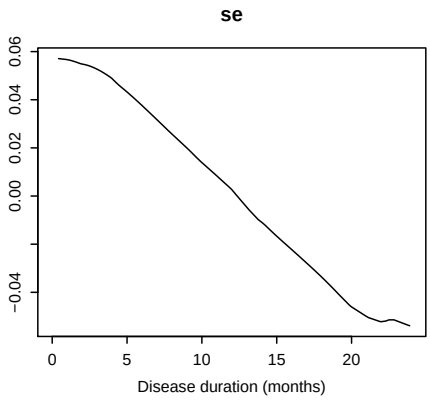

imd07_q4
0.3
0.2
0.1
0.0
−0.1
0
5
10
15
20
Disease duration (months)

Figure 4: Plots of estimates of functional coefficient versus disease durations – HAQ progression data with selected covariates for the logit of $Y = 3$ vs $Y = 1$.




age
−0.001
0.000
0.001
0.002
0
5
10
15
20
Disease duration (months)


se
−0.04
0.00
0.02
0.04
0.06
0
5
10
15
20
Disease duration (months)

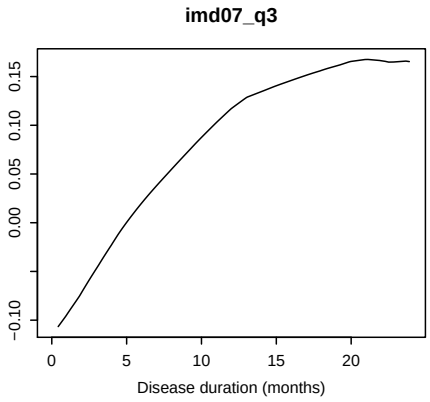

imd07_q3
−0.10
0.00
0.05
0.10
0.15
0
5
10
15
20
Disease duration (months)

Table 8: HAQ progression data – selected covariates for the logit of $Y = 2$ vs $Y = 1$ and their coefficient estimates (standard errors): either constant or varying at selected disease durations

Variable with constant coefficient	Constant				
ra	0.0907 (0.0513)	-	-	-	-
gender	0.1444 (0.0103)	-	-	-	-
smoker_now	0.0925 (0.0021)	-	-	-	-
Variable with Functional coefficient	Month(s)				
	1	3	6	12	24
haq0	0.1472 (0.0411)	0.0835 (0.0190)	0.0439 (0.0166)	0.0893 (0.0226)	0.5764 (0.1534)
swollen_51jt	-0.0112 (0.0040)	-0.0045 (0.0016)	0.0012 (0.0003)	0.0012 (0.0004)	-0.0278 (0.0068)
tend_51jt	-0.0073 (0.0026)	-0.0044 (0.0010)	-0.0004 (0.0001)	0.0067 (0.0018)	0.0133 (0.0050)
dmard	-0.0002 (0.0001)	-0.0001 (0.0001)	0.0001 (0.0001)	0.0002 (0.0001)	0.0002 (0.0001)
age	-0.0017 (0.0005)	-0.0010 (0.0002)	-0.0004 (0.0001)	0.0000 (0.0001)	-0.0002 (0.0001)
se	0.0778 (0.0275)	0.0536 (0.0214)	0.0382 (0.0149)	0.0058 (0.0018)	-0.1035 (0.0329)
smoke_pre	0.1067 (0.0273)	0.0796 (0.0194)	0.0304 (0.0093)	-0.0319 (0.0066)	-0.1732 (0.0574)
imd07_q3	0.2911 (0.0860)	0.2412 (0.0887)	0.1944 (0.0464)	0.0830 (0.0321)	-0.1844 (0.0531)
imd07_q4	0.3086 (0.1005)	0.2896 (0.0934)	0.2637 (0.0877)	0.1740 (0.0672)	-0.1755 (0.0675)

selection and model specification by a penalised likelihood approach. Based on the ideas of penalization on deviation, kernel smoothing and quadratic function approximation, our method selects significant predictors, determines whether each selected predictor has a constant or functional coefficient and estimates their coefficients at the same time.

Table 9: HAQ progression data – selected covariates for the logit of $Y = 3$ vs $Y = 1$ and their coefficient estimates (standard errors): either constant or varying at selected disease durations

Variable with constant coefficient	Constant				
haq0	0.0515 (0.0004)	-	-	-	-
rf	0.0771 (0.0008)	-	-	-	-
Variable with	Month(s)				
Functional coefficient	1	3	6	12	24
intercept	-0.1585 (0.0433)	-0.1493 (0.0387)	-0.1379 (0.0521)	-0.0976 (0.0336)	0.0247 (0.0069)
swollen_51jt	0.0050 (0.0019)	0.0043 (0.0013)	0.0033 (0.0007)	0.0038 (0.0013)	0.0098 (0.0033)
tend_51jt	-0.0055 (0.0021)	-0.0038 (0.0011)	-0.0013 (0.0003)	0.0021 (0.0007)	0.0147 (0.0056)
dmard	0.0006 (0.0002)	0.0004 (0.0002)	0.0001 (0.0001)	-0.0002 (0.0002)	-0.0001 (0.0001)
age	0.0026 (0.0009)	0.0023 (0.0006)	0.0020 (0.0004)	0.0007 (0.0002)	-0.0010 (0.0003)
se	0.0567 (0.0182)	0.0525 (0.0160)	0.0377 (0.0149)	0.0027 (0.0008)	-0.0541 (0.0208)
imd07_q3	-0.0948 (0.0193)	-0.0453 (0.0160)	0.0197 (0.0075)	0.1171 (0.0440)	0.1649 (0.0379)

Another attractive feature of the proposed method is that it allows the number of potential covariates to increase with the sample size. With rapid development of laboratory medicine, more potential prognostic markers, including clinical and demographic features, environmental factors, serological factors, genetic factors, epigenetic factors and their interactions, are considered as candidates to predict future disease outcome or response to treatment in stratified medicine and the number of potential predictors could be very large.

Table 10: HAQ progression data – comparison of correct classification rates among models

Estimated conditional probability $\geq 80\%$			
Model	Total classification No.	Correct classification No.	Correct classification rate
Model 1	61	52	85.2%
Model 2	77	51	66.2%
Model 3	45	37	82.1%
Model 4	34	22	71.0%
Model 5	81	50	61.7%
Estimated conditional probability $\geq 70\%$			
Model	Total classification No.	Correct classification No.	Correct classification rate
Model 1	120	93	77.5%
Model 2	138	88	63.7%
Model 3	142	105	73.9%
Model 4	70	46	65.7%
Model 5	143	90	62.9%

"Total classification No." is the number of classifiable subjects whose estimated maximum group-membership probability is higher or equal to 80% *or* 70%. *"Correct classification No." is the number of correctly classified subjects. "Correct classification rate" is the ratio between "Correct classification No." and "Total classification No.".*

This study focuses on nonparametric prognostic classification modelling and future work would be focussing on treatment-specific consideration in stratified medicine and methods to predict response to treatment.

9 Proofs of the theoretical results in Chapter 4.1

In Chapter 9.1, we give some assumptions which are needed to prove the asymptotic theory in Chapter 4.1. In Chapter 9.2 and Chapter 9.3, we provide the proofs of the main theoretical results and some auxiliary results, respectively.

9.1 Assumptions

Recall that

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

Let $\mu_k=\int u^k K(u)du$ and $\nu_k=\int u^k K^2(u)du$ for $k=0,1,2,\cdots$, $\Lambda(u)=f_U(u)\mathsf{diag}\big(1,\mu_2\big)$,

$$V_1(u)=\mathsf{E}\Big\{q_1^2\big[\sum_{j=1}^{d_n}a_j(U)x_j,y\big]XX^{\mathrm{T}}|U=u\Big\}$$

and

$$V_2(u)=\mathsf{E}\Big\{q_2\big[\sum_{j=1}^{d_n}a_j(U)x_j,y\big]XX^{\mathrm{T}}|U=u\Big\},$$

where $f_U(\cdot)$ is the marginal density function of U. Define

$$a_{n1}=\max_{1\le j\le d_n(1)}\dot{p}_{\lambda_{1j}}\big(\|\widetilde{\boldsymbol{\beta}}_j\|\big),\quad a_{n2}=\max_{1\le j\le d_n(2)}\dot{p}_{\lambda_{2j}}\big(\|\widetilde{\boldsymbol{\alpha}}_j\|\big).$$

and

$$b_{n1} = \min_{d_n(1)+1 \le j \le d_n} \dot{p}_{\lambda_{1j}}\big(\|\widetilde{\boldsymbol{\beta}}_j\|\big), \quad b_{n2} = \min_{d_n(2)+1 \le j \le d_n} \dot{p}_{\lambda_{2j}}\big(\|\widetilde{\boldsymbol{\alpha}}_j\|\big).$$

Let $\alpha_n \propto \beta_n$ denote $c_1\beta_n \le \alpha_n \le c_2\beta_n$ for $0 < c_1 \le c_2 < \infty$. We next introduce some regularity conditions which have been used in Chapter 4.1 to establish the asymptotic theory. Some of them might be not the weakest possible conditions.

Assumption A1. The kernel function $K(\cdot)$ is a continuous and symmetric probability density function with a compact support.

Assumption A2. *(i)* Let $\mathsf{E}\Big\{q_1^2\big[\sum_{j=1}^{d_n} a_j(U)x_j, y\big]|U = u\Big\}$ be continuous for $u \in [0,\ 1]$ and

$$\mathsf{E}\Big\{\big|q_1\big[\sum_{j=1}^{d_n} a_j(U)x_j, y\big]\big|^{2+\delta}\Big\} < \infty$$

for some $\delta > 0$.

(ii) Let $q_2(s, y) < 0$ for $s \in \mathbb{R}$ and y in the range of the response variable. Furthermore, $\mathsf{E}\Big\{\big|q_2\big[\sum_{j=1}^{d_n} a_j(U)x_j, y\big]\big|^{2+\delta}\Big\} < \infty$ and there exists a $M(X, U, y) > 0$ such that

$$\Big|q_2(s_2(X,U), y) - q_2(s_1(X,U), y)\Big| \le M(X,U,y)\Big|s_2(X,U) - s_1(X,U)\Big|$$

and

$$\max_{i,j,k} \sup_u \mathsf{E}\Big[\big|x_i x_j x_k\big| M(X,U,y)\Big|U = u\Big] < \infty.$$

(iii) Let $V_1(u)$ and $V_2(u)$ be continuous for $u \in [0,\ 1]$, and $-\Lambda(u) \otimes V_2(u)$ be positive definite for any $u \in [0,\ 1]$ with eigenvalues bounded away from zero and infinity, where $\otimes$ denotes the Kronecker product.

Assumption A3. 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 A4. The functional coefficients, $a_j(\cdot)$ have continuous second-order derivative for $j = 1, \cdots, d_n$.

Assumption A5. Let the bandwidth $h \propto n^{-1/3}$ and the number of the covariates $d_n = o(n^{4/15} \log^{-1/3} n)$.

Assumption A6. *(i)* The penalty functions, $p_{\lambda_{kj}}(\cdot)$, are positive and nondecreasing on $(0, \infty)$ and have the first-order derivatives denoted by $\dot{p}_{\lambda_{kj}}(\cdot)$ for $k = 1, 2$ and $j = 1, 2, \cdots, d_n$. In addition, $\dot{p}_{\lambda_{kj}}(z) \geq 0$ if $z \geq 0$.

(ii) Let $a_{n1} = o_P(\gamma_n n^{3/2} h/\sqrt{d_n})$ and $a_{n2} = o_P(\gamma_n n^{3/2}/\sqrt{d_n})$, where $\gamma_n = \sqrt{\frac{d_n}{nh}}$.

(iii) Let $\gamma_n n^{3/2}/b_{n1} = o_P(1)$ and $\gamma_n n^{3/2}/b_{n2} = o_P(1)$.

Assumption A7. Let $a_{n1} = o_P(\sqrt{n/d_n})$ and $a_{n2} = o_P(\sqrt{n/d_n})$.

Remark A.1. The above assumptions are mild and justifiable. Assumption A1 is a mild condition on the kernel function and the compact support restriction can be relaxed at the cost of more tedious proofs.

Assumption A2 imposes some smoothness and moment conditions on $q_1(\cdot,\cdot)$ and $q_2(\cdot,\cdot)$, which are commonly used in local maximum likelihood estimation (see, for example, Cai *et al* 2000, Li and Liang 2008). Assumptions A3 and A4 provide some smoothness conditions on the density function of U and the functional coefficients $a_j(\cdot)$, which are necessary when the local linear approach is applied (see, for example, Fan and Gijbels 1996). In Assumption A5, we let the bandwidth chosen as the optimal rate, and allow that the dimension of the covariates diverges with a polynomial rate. Assumption A6 imposes some restrictions on the penalty functions and the tuning parameters λ_{1j} and λ_{2j}. We will later show in Chapter 9.3 that the SCAD and LASSO penalty functions would satisfy these conditions with mild restrictions on the tuning parameters. Some additional restrictions on the penalty term in Assumption A7 are mainly used to establish the oracle property in Theorem 4.2. However, if we are only interested on the oracle property for the nonparametric estimation in Theorem 4.2 and can prove (4.1) and (4.2) in Proposition 4.1 for the penalised local maximum estimates $\widehat{\mathbf{a}}(\cdot)$ and $\widehat{\mathbf{b}}(\cdot)$, the conditions in Assumption A7 can be relaxed to $a_{n1} = o_P(n/\sqrt{d_n h})$ and $a_{n2} = o_P(n/\sqrt{d_n h})$.

9.2 Proofs of the main results

We now provide the detailed proofs of the asymptotic results stated in Chapter 4.1. Define

$$\mathcal{L}_{nu}(\mathbf{a},\mathbf{b}) = \sum_{i=1}^{n} \ell\Big(g^{-1}\Big\{\sum_{j=1}^{d_n}\big[\alpha_j+\beta_j(U_i-u)\big]x_{ij}\Big\},\ y_i\Big)K_h(U_i-u) \quad (9.1)$$

for $u \in [0,1]$. Let $\widetilde{\mathbf{a}}(u)$ and $\widetilde{\mathbf{b}}(u)$ be the maximiser to $\mathcal{L}_{nu}(\mathbf{a},\mathbf{b})$, the local maximum likelihood estimators of $\mathbf{a}(u)$ and $\mathbf{b}(u)$, where

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

We first give the proof of the uniform consistency results in Proposition 4.1.

Proof of Proposition 4.1. To prove (4.1) and (4.2), it suffices to show that

$$\sup_{u\in[0,1]} \big\|\widetilde{\mathbf{a}}(u) - \mathbf{a}(u)\big\| = O_P\Big(\sqrt{\frac{d_n \log n}{nh}}\Big) \tag{9.2}$$

and

$$\sup_{u\in[0,1]} \big\|h[\widetilde{\mathbf{b}}(u) - \mathbf{b}(u)]\big\| = O_P\Big(\sqrt{\frac{d_n \log n}{nh}}\Big). \tag{9.3}$$

In order to prove (9.2) and (9.3), we first prove the result that uniformly for $u \in [0,1]$,

$$\begin{pmatrix} \widetilde{\mathbf{a}}(u) - \mathbf{a}(u) \\ h[\widetilde{\mathbf{b}}(u) - \mathbf{b}(u)] \end{pmatrix} = -\ddot{\mathcal{L}}^{+}_{nu}(\mathbf{a}(u),\mathbf{b}(u))\dot{\mathcal{L}}_{nu}(\mathbf{a}(u),\mathbf{b}(u))(1+o_P(1)), \tag{9.4}$$

where $\mathbb{A}^+$ is the Moore-Penrose inverse matrix of $\mathbb{A}$,

$$\begin{aligned}
\dot{\mathcal{L}}_{nu}(\mathbf{a},\mathbf{b}) =& \sum_{i=1}^{n} q_1\Big(\Big\{\sum_{j=1}^{d_n}\big[\alpha_j+\beta_j(U_i-u)\big]x_{ij}\Big\},\ y_i\Big)K_h(U_i-u)\\
&\begin{pmatrix} 1 \\ \frac{U_i-u}{h} \end{pmatrix}\otimes X_i,\\
\ddot{\mathcal{L}}_{nu}(\mathbf{a},\mathbf{b}) =& \sum_{i=1}^{n} q_2\Big(\Big\{\sum_{j=1}^{d_n}\big[\alpha_j+\beta_j(U_i-u)\big]x_{ij}\Big\},\ y_i\Big)K_h(U_i-u)\\
&\begin{bmatrix} 1 & \frac{U_i-u}{h} \\ \frac{U_i-u}{h} & \frac{(U_i-u)^2}{h^2} \end{bmatrix}\otimes X_iX_i^{\mathrm{T}}.
\end{aligned}$$

By Taylor's expansion for $\dot{\mathcal{L}}_{nu}(\widetilde{\mathbf{a}}(u),\widetilde{\mathbf{b}}(u))$ at $(\mathbf{a}(u),\mathbf{b}(u))$, we have

$$\begin{aligned}
0 =& \dot{\mathcal{L}}_{nu}(\widetilde{\mathbf{a}}(u),\widetilde{\mathbf{b}}(u))\\
=& \dot{\mathcal{L}}_{nu}(\mathbf{a}(u),\mathbf{b}(u)) + \ddot{\mathcal{L}}_{nu}(\mathbf{a}_*(u),\mathbf{b}_*(u))\begin{pmatrix} \widetilde{\mathbf{a}}(u)-\mathbf{a}(u) \\ h[\widetilde{\mathbf{b}}(u)-\mathbf{b}(u)] \end{pmatrix},
\end{aligned}$$

where $\mathbf{a}_*(u)$ lies between $\mathbf{a}(u)$ and $\widetilde{\mathbf{a}}(u)$, and $\mathbf{b}_*(u)$ lies between $\mathbf{b}(u)$ and $\widetilde{\mathbf{b}}(u)$. As in the proof of Lemma A.2 in Zhang *et al* (2012), we may show that

$$\ddot{\mathcal{L}}_{nu}(\mathbf{a}_*(u),\mathbf{b}_*(u)) = \ddot{\mathcal{L}}_{nu}(\mathbf{a}(u),\mathbf{b}(u))(1+o_P(1))$$

uniformly for $u\in[0,1]$. Then, using the Convex Lemma (c.f., Pollard 1991), we can prove (9.4) uniformly for $u\in[0,1]$.

By Assumptions A1–A5 in Chapter 9.1, we have, uniformly for

$u \in [0,1]$,

$$\frac{1}{n}\ddot{\mathcal{L}}_{nu}(\mathbf{a}(u),\mathbf{b}(u)) = \Lambda(u) \otimes V_2(u)(1+o_P(1)). \tag{9.5}$$

The detailed proof of (9.5) will be given later in Chapter 9.3.

Let $R_i = \sum_{j=1}^{d_n} a_j(U_i)x_{ij}$. We next consider $\dot{\mathcal{L}}_{nu}(\mathbf{a}(u),\mathbf{b}(u))$. Let

$$\dot{\mathcal{L}}_n(u) = \sum_{i=1}^{n} q_1(R_i,y_i)K_h(U_i-u)\begin{pmatrix} 1 \\ \frac{U_i-u}{h} \end{pmatrix} \otimes X_i =: \big[S_{n,1}(u),\cdots,S_{n,2d_n}(u)\big]^{\mathrm{T}}.$$

Observe that

$$\begin{aligned}
\dot{\mathcal{L}}_{nu}(\mathbf{a}(u),\mathbf{b}(u)) =& \dot{\mathcal{L}}_n(u) + \big[\dot{\mathcal{L}}_{nu}(\mathbf{a}(u),\mathbf{b}(u)) - \dot{\mathcal{L}}_n(u)\big] \\
=:& \dot{\mathcal{L}}_n(u) + \big[T_{n,1}(u),\cdots,T_{n,2d_n}(u)\big]^{\mathrm{T}} \\
=& \big[S_{n,1}(u)+T_{n,1}(u),\cdots,S_{n,2d_n}(u)+T_{n,2d_n}(u)\big]^{\mathrm{T}}.
\end{aligned} \tag{9.6}$$

By Assumptions A1–A3, A5 and similarly to the proof of Theorem B in Mack and Silverman (1982), we can show that

$$\sup_{u\in[0,1]} \frac{1}{n}\big|S_{n,k}(u)\big| = O_P\Big(\sqrt{\frac{\log n}{nh}}\Big), \quad k=1,\cdots,2d_n. \tag{9.7}$$

By Taylor's expansion for $q_1(r,y)$ with respect to r,

$$\sup_{u\in[0,1]} \frac{1}{n}\big|T_{n,k}(u)\big| = O_P(h^2), \quad k=1,\cdots,2d_n. \tag{9.8}$$

Let $\lambda_s(u)$ and $\lambda_l(u)$ be the smallest and largest eigenvalues of

$-\Lambda(u) \otimes V_2(u)$, respectively, and

$$\lambda_s = \inf_{u\in[0,1]} \lambda_s(u), \quad \lambda_l = \sup_{u\in[0,1]} \lambda_l(u).$$

By Assumption A2(iii), it is easy to show that $0 < \lambda_s \le \lambda_l < \infty$, which implies that the largest eigenvalue of $\big[-\Lambda(u) \otimes V_2(u)\big]^+$ is bounded, and

$$\max_{\|\mathbf{z}\|=1} \big\| \big[\Lambda(u) \otimes V_2(u)\big]^+ \mathbf{z} \big\| < \infty \tag{9.9}$$

uniformly for $u \in [0,1]$. Hence, by (9.4)–(9.9) and noting that $h^2 = o\big(\sqrt{\frac{\log n}{nh}}\big)$ by Assumption A5, we can prove (9.2) and (9.3).

□

Proof of Proposition 4.2. Let

$$\mathcal{A}_0 = \big[\mathbf{a}^{\mathrm{T}}(U_1),\ \cdots,\ \mathbf{a}^{\mathrm{T}}(U_n)\big]^{\mathrm{T}}, \quad \mathcal{B}_0 = \big[\mathbf{b}^{\mathrm{T}}(U_1),\ \cdots,\ \mathbf{b}^{\mathrm{T}}(U_n)\big]^{\mathrm{T}},$$

and

$$\mathcal{U} = \big[\mathbf{u}^{\mathrm{T}}(1), \cdots, \mathbf{u}^{\mathrm{T}}(n)\big]^{\mathrm{T}}, \quad \mathcal{V} = \big[\mathbf{v}^{\mathrm{T}}(1), \cdots, \mathbf{v}^{\mathrm{T}}(n)\big]^{\mathrm{T}},$$

where both $\mathbf{u}(k)$ and $\mathbf{v}(k)$ are column vectors with dimension d_n for $k = 1, \cdots, n$. Define

$$\Omega(C) = \big\{(\mathcal{U}, \mathcal{V}) :\ \|\mathcal{U}\|^2 = nC, \|\mathcal{V}\|^2 = nC\big\},$$

where C is a positive constant.

For $(\mathcal{U}, \mathcal{V}) \in \Omega(C)$, observe that

$$Q_{n*}\big(\mathcal{A}_0 + \gamma_n \mathcal{U}, \mathcal{B}_0 + \gamma_n \mathcal{V}/h\big) - Q_{n*}(\mathcal{A}_0, \mathcal{B}_0) = I_{n1} + I_{n2} + I_{n3}, \tag{9.10}$$

where

$$
\begin{aligned}
I_{n1} =& \big[\mathcal{L}_{n*}\big(\mathcal{A}_0 + \gamma_n \mathcal{U}, \mathcal{B}_0 + \gamma_n \mathcal{V}/h\big) - \mathcal{L}_{n*}(\mathcal{A}_0, \mathcal{B}_0)\big], \\
I_{n2} =& \sum_{j=1}^{d_n} \big[\mathcal{P}_{1n,j}(\|\boldsymbol{\beta}_{j0}\|) - \mathcal{P}_{1n,j}(\|\boldsymbol{\beta}_{j0} + \gamma_n \mathbf{v}_j/h\|)\big], \\
I_{n3} =& \sum_{j=1}^{d_n} \big[\mathcal{P}_{2n,j}(\|\boldsymbol{\alpha}_{j0}\|) - \mathcal{P}_{2n,j}(\|\boldsymbol{\alpha}_{j0} + \gamma_n \mathbf{u}_j\|)\big],
\end{aligned}
$$

in which γ_n is defined in Assumption A6(ii), $\boldsymbol{\alpha}_{j0} = \big[a_j(U_1), \cdots, a_j(U_n)\big]^{\mathrm{T}}$, $\boldsymbol{\beta}_{j0} = \big[\dot{a}_j(U_1), \cdots, \dot{a}_j(U_n)\big]^{\mathrm{T}}$, $\mathbf{u}_j = \big[u_j(1), \cdots, u_j(n)\big]^{\mathrm{T}}$, $\mathbf{v}_j = \big[v_j(1), \cdots, v_j(n)\big]^{\mathrm{T}}$, $u_j(k)$ and $v_j(k)$ are the j-th component of vectors $\mathbf{u}(k)$ and $\mathbf{v}(k)$, respectively.

We first consider I_{n1}. By the definition of $\mathcal{L}_{n*}(\cdot, \cdot)$ in Chapter 3.2, we have

$$
\begin{aligned}
I_{n1} =& \mathcal{L}_{n*}\big(\mathcal{A}_0 + \gamma_n \mathcal{U}, \mathcal{B}_0 + \gamma_n \mathcal{V}/h\big) - \mathcal{L}_{n*}\big(\mathcal{A}_0, \mathcal{B}_0\big), \qquad (9.11) \\
\overset{P}{\sim}& \gamma_n \big(\mathcal{U}^{\mathrm{T}}, \mathcal{V}^{\mathrm{T}}\big) \dot{\mathcal{L}}_n(\mathcal{A}_0, \mathcal{B}_0) + \frac{1}{2}\gamma_n^2 \big(\mathcal{U}^{\mathrm{T}}, \mathcal{V}^{\mathrm{T}}\big) \ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n) \begin{pmatrix} \mathcal{U} \\ \mathcal{V} \end{pmatrix},
\end{aligned}
$$

where $a_n \overset{P}{\sim} b_n$ denotes that $a_n = b_n(1 + o_P(1))$. The detailed proof of (9.11) will be provided in Chapter 9.3 below.

We define

$$
\begin{aligned}
I_{n4} =& \gamma_n \big(\mathcal{U}^{\mathrm{T}}, \mathcal{V}^{\mathrm{T}}\big) \dot{\mathcal{L}}_n(\mathcal{A}_0, \mathcal{B}_0), \\
I_{n5} =& \frac{1}{2}\gamma_n^2 \big(\mathcal{U}^{\mathrm{T}}, \mathcal{V}^{\mathrm{T}}\big) \ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n) \begin{pmatrix} \mathcal{U} \\ \mathcal{V} \end{pmatrix}.
\end{aligned}
$$

Using Assumptions A1–A5 and Cauchy-Schwarz inequality, and by some elementary but tedious calculations, we can show that

$$I_{n4} = O_P(\gamma_n^2 n^{3/2}) \cdot \big(\|\mathcal{U}\| + \|\mathcal{V}\|\big). \tag{9.12}$$

The detailed proof of (9.12) will be also given in Chapter 9.3 below.

For I_{n5}, note that

$$\begin{aligned} I_{n5} =& \frac{1}{2}\gamma_n^2 (\mathcal{U}^{\mathrm{T}}, \mathcal{V}^{\mathrm{T}}) \Big[\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n) - \ddot{\mathcal{L}}_n(\mathcal{A}_0, \mathcal{B}_0) \Big] \begin{pmatrix} \mathcal{U} \\ \mathcal{V} \end{pmatrix} \\ & + \frac{1}{2}\gamma_n^2 (\mathcal{U}^{\mathrm{T}}, \mathcal{V}^{\mathrm{T}}) \ddot{\mathcal{L}}_n(\mathcal{A}_0, \mathcal{B}_0) \begin{pmatrix} \mathcal{U} \\ \mathcal{V} \end{pmatrix} \\ =:& I_{n6} + I_{n7}. \end{aligned} \tag{9.13}$$

Recalling that $\lambda_s(u)$ is the smallest eigenvalue for $-\Lambda(u) \otimes V_2(u)$, by Assumption A2(iii), we have $\lambda_s = \inf_{u\in[0,1]} \lambda_s(u) > 0$. Then, following the proof of (9.5) in Chapter 9.3, we can show that

$$I_{n7} \le -\lambda_s \gamma_n^2 n \cdot \big(\|\mathcal{U}\|^2 + \|\mathcal{V}\|^2\big) < 0. \tag{9.14}$$

By Assumptions A2(ii) and A5, and using Proposition 4.1, we can

prove that

$$\begin{aligned} I_{n6} =& O_P(d_n\gamma_n^3 n\sqrt{\log n}) \cdot \max_{i,j,k} \sup_u \mathsf{E}\Big[|x_i x_j x_k| M(X,U,y)\Big| U=u\Big] \\ & \cdot \Big(\|\mathcal{U}\|^2 + \|\mathcal{V}\|^2\Big) \\ =& O_P(d_n\gamma_n^3 n\sqrt{\log n}) \cdot \Big(\|\mathcal{U}\|^2 + \|\mathcal{V}\|^2\Big) \\ =& o_P(\gamma_n^2 n) \cdot \Big(\|\mathcal{U}\|^2 + \|\mathcal{V}\|^2\Big), \end{aligned}$$

which, together with (9.11)–(9.14), implies that I_{n7} is the leading term of I_{n1}. Hence, when n is sufficiently large, by taking C large enough, we have

$$I_{n1} \overset{P}{\sim} \frac{1}{2}\gamma_n^2 (\mathcal{U}^{\mathrm{T}}, \mathcal{V}^{\mathrm{T}}) \ddot{\mathcal{L}}_n(\mathcal{A}_0, \mathcal{B}_0) \begin{pmatrix} \mathcal{U} \\ \mathcal{V} \end{pmatrix}. \tag{9.15}$$

We next consider I_{n2}. By the definition of $\mathcal{P}_{1n,j}(\cdot)$ and Assumption

A6(ii), we have

$$
\begin{aligned}
I_{n2} =& \sum_{j=1}^{d_n} \big[\mathcal{P}_{1n,j}(\boldsymbol{\beta}_{j0}) - \mathcal{P}_{1n,j}(\boldsymbol{\beta}_{j0} + \gamma_n \mathbf{v}_j/h)\big] \\
=& \sum_{j=1}^{d_n} \dot{p}_{\lambda_{1j}}\big(\|\widetilde{\boldsymbol{\beta}}_j\|\big)\big(\|\boldsymbol{\beta}_{j0}\| - \|\boldsymbol{\beta}_{j0} + \gamma_n \mathbf{v}_j/h\|\big) \\
\le& \sum_{j=1}^{d_n(1)} \dot{p}_{\lambda_{1j}}\big(\|\widetilde{\boldsymbol{\beta}}_j\|\big)\big(\|\boldsymbol{\beta}_{j0}\| - \|\boldsymbol{\beta}_{j0} + \gamma_n \mathbf{v}_j/h\|\big) \\
& - \sum_{j=d_n(1)+1}^{d_n} \dot{p}_{\lambda_{1j}}\big(\|\widetilde{\boldsymbol{\beta}}_j\|\big)\|\boldsymbol{\beta}_{j0} + \gamma_n \mathbf{v}_j/h\| \\
=& O_P\big(\sqrt{d_n(1)}\gamma_n a_{n1}/h\big) \cdot \|\mathcal{V}\| - \sum_{j=d_n(1)+1}^{d_n} \dot{p}_{\lambda_{1j}}\big(\|\widetilde{\boldsymbol{\beta}}_j\|\big)\|\boldsymbol{\beta}_{j0} + \gamma_n \mathbf{v}_j/h\| \\
=& o_P(\gamma_n^2 n) \cdot \|\mathcal{V}\|^2 - \sum_{j=d_n(1)+1}^{d_n} \dot{p}_{\lambda_{1j}}\big(\|\widetilde{\boldsymbol{\beta}}_j\|\big)\|\boldsymbol{\beta}_{j0} + \gamma_n \mathbf{v}_j/h\|. \qquad (9.16)
\end{aligned}
$$

Similarly, by the definition of $\mathcal{P}_{2n,j}(\cdot)$ and Assumption A6(ii) again, we also have

$$
\begin{aligned}
I_{n3} =& O_P\Big(\sqrt{d_n(2)}\gamma_n a_{n2}\Big) \cdot \|\mathcal{U}\| - \sum_{j=d_n(2)+1}^{d_n} \dot{p}_{\lambda_{2j}}\big(\|\widetilde{\boldsymbol{\alpha}}_j\|\big)\|\boldsymbol{\alpha}_{j0} + \gamma_n \mathbf{u}_j\| \\
=& o_P(\gamma_n^2 n) \cdot \|\mathcal{U}\|^2 - \sum_{j=d_n(2)+1}^{d_n} \dot{p}_{\lambda_{2j}}\big(\|\widetilde{\boldsymbol{\alpha}}_j\|\big)\|\boldsymbol{\alpha}_{j0} + \gamma_n \mathbf{u}_j\|. \qquad (9.17)
\end{aligned}
$$

Hence, by (9.10) and (9.15)–(9.17), we can prove that the leading term of $I_{n1} + I_{n2} + I_{n3}$ is negative in probability, which indicates that

for any $\epsilon > 0$, there exists a sufficiently large $C > 0$ such that

$$\mathsf{P}\left\{\sup_{(\mathcal{U},\mathcal{V})\in\Omega(C)} Q_{n*}\big(\mathcal{A}_0+\gamma_n\mathcal{U},\mathcal{B}_0+\gamma_n\mathcal{V}/h\big) < Q_{n*}(\mathcal{A}_0,\mathcal{B}_0)\right\} \geq 1-\epsilon \tag{9.18}$$

for large n, which implies that (4.3) and (4.4) holds. □

Proof of Theorem 4.1. To prove (4.5), it is equivalent to show

$$\mathsf{P}\left(\max_{d_n(2)+1\leq j\leq d_n} \|\widehat{\boldsymbol{\alpha}}_j\| \neq 0\right) \to 0 \tag{9.19}$$

as n tends to infinity. As a consequence of the Karush-Kuhn-Tucker conditions, for any $d_n(2)+1 \leq j \leq d_n$ such that $\|\widehat{\boldsymbol{\alpha}}_j\| \neq 0$, we must have

$$\frac{\partial \mathcal{L}_{n*}(\mathcal{A},\mathcal{B})}{\partial \boldsymbol{\alpha}_j} = \dot{p}_{\lambda_{2j}}(\|\widetilde{\boldsymbol{\alpha}}_j\|)\frac{\boldsymbol{\alpha}_j}{\|\boldsymbol{\alpha}_j\|} \tag{9.20}$$

when $\boldsymbol{\alpha}_j = \widehat{\boldsymbol{\alpha}}_j$. It is easy to see that the Euclidean norm of the right hand side of equation (9.20) is larger than b_{n2}, which is defined in Chapter 9.1 and is independent of j. Note that the convergence rates in Proposition 4.2 hold for both the local maximum likelihood estimation and penalised maximum likelihood estimation. Following the proof of (9.9), we may show that the Euclidean norm of the left hand side of (9.20) is bounded by $O_P(n^{3/2}\gamma_n)$ uniformly for $d_n(2)+1 \leq j \leq d_n$. Assumption A6(iii) indicates that the probability for (9.20) holds for at least one $d_n(2)+1 \leq j \leq d_n$ is zero as n tends to infinity. Hence, we can prove that (9.19) holds.

Similarly, the proof of (4.6) is equivalent to the proof of

$$\mathsf{P}\left(\max_{d_n(1)+1\le j\le d_n}\|\widehat{\boldsymbol{\beta}}_j\|\neq 0\right)\to 0 \tag{9.21}$$

as n tends to infinity. Applying the Karush-Kuhn-Tucker conditions, for any $d_n(1)+1\le j\le d_n$ such that $\|\widehat{\boldsymbol{\beta}}_j\|\neq 0$, we must have

$$\frac{\partial\mathcal{L}_{n*}(\mathcal{A},\mathcal{B})}{\partial\boldsymbol{\beta}_j}=\dot{p}_{\lambda_{2j}}(\|\widetilde{\boldsymbol{\beta}}_j\|)\frac{\boldsymbol{\beta}_j}{\|\boldsymbol{\beta}_j\|} \tag{9.22}$$

when $\boldsymbol{\beta}_j=\widehat{\boldsymbol{\beta}}_j$. Using the argument analogous to the proof of (9.19) and Assumption A6(iii), we can also prove that (9.21) holds. We then complete the proof of Theorem 4.1. □

Proof of Theorem 4.2. The proof is similar to the proof of Theorem 2 in Wang and Xia (2009) with some modifications. Let $X_i^*=\big[x_{i1},\cdots,x_{id_n(1)}\big]^{\mathrm{T}}$,

$$\begin{aligned}\dot{\mathcal{L}}_{nu}^*(\mathbf{a},\mathbf{b})=&\sum_{i=1}^n q_1\Big(\Big\{\sum_{j=1}^{d_n}\big[\alpha_j+\beta_j(U_i-u)\big]x_{ij}\Big\},\ y_i\Big)K_h(U_i-u)\\ &\begin{pmatrix}1\\ \frac{U_i-u}{h}\end{pmatrix}\otimes X_i^*,\\ \ddot{\mathcal{L}}_{nu}^*(\mathbf{a},\mathbf{b})=&\sum_{i=1}^n q_2\Big(\Big\{\sum_{j=1}^{d_n}\big[\alpha_j+\beta_j(U_i-u)\big]x_{ij}\Big\},\ y_i\Big)K_h(U_i-u)\\ &\begin{bmatrix}1 & \frac{U_i-u}{h}\\ \frac{U_i-u}{h} & \frac{(U_i-u)^2}{h^2}\end{bmatrix}\otimes X_i^*(X_i^*)^{\mathrm{T}}.\end{aligned}$$

For $i = 1, \cdots, n$, denote

$$\widehat{\mathbf{a}}_1(U_i) = \big[\widehat{a}_1(U_i), \cdots, \widehat{a}_{d_n(1)}(U_i)\big]^{\mathrm{T}},$$
$$\widehat{\mathbf{a}}_o(U_i) = \big[\widehat{a}_{1o}(U_i), \cdots, \widehat{a}_{d_n(1)o}(U_i)\big]^{\mathrm{T}},$$

and let $\widehat{\mathbf{b}}_1(U_i)$ and $\widehat{\mathbf{b}}_o(U_i)$ be the penalised and oracle local maximum estimates of $\big[\dot{a}_1(U_i), \cdots, \dot{a}_{d_n(1)}(U_i)\big]^{\mathrm{T}}$, respectively.

Following the proof of (9.11) in Chapter 9.3, we can show that the oracle estimates satisfy the following equation:

$$\mathbf{0} = \dot{\mathcal{L}}^*_{nU_i}\big(\mathbf{a}(U_i), \mathbf{b}(U_i)\big) + \ddot{\mathcal{L}}^*_{nU_i}\big(\mathbf{a}(U_i), \mathbf{b}(U_i)\big) \begin{bmatrix} \widehat{\mathbf{a}}_o(U_i) - \mathbf{a}_1(U_i) \\ \widehat{\mathbf{b}}_o(U_i) - \mathbf{b}_1(U_i) \end{bmatrix} \tag{9.23}$$

uniformly for $1 \le i \le n$, where $\mathbf{a}_1(u)$ and $\mathbf{b}_1(u)$ are the sub-vectors consisting of the first $d_n(1)$ elements of $\mathbf{a}(u)$ and $\mathbf{b}(u)$, respectively.

Following the proof of Theorem 4.1, we can also show that the penalised estimates satisfy the following equation:

$$\begin{aligned} \mathbf{0} =& \dot{\mathcal{L}}^*_{nU_i}\big(\mathbf{a}(U_i), \mathbf{b}(U_i)\big) + \ddot{\mathcal{L}}^*_{nU_i}\big(\mathbf{a}(U_i), \mathbf{b}(U_i)\big) \begin{bmatrix} \widehat{\mathbf{a}}_1(U_i) - \mathbf{a}_1(U_i) \\ \widehat{\mathbf{b}}_1(U_i) - \mathbf{b}_1(U_i) \end{bmatrix} \\ & - \Big[\mathbf{P}^{\mathrm{T}}_{\mathbf{a}}(U_i), \mathbf{P}^{\mathrm{T}}_{\mathbf{b}}(U_i)\Big]^{\mathrm{T}} \end{aligned} \tag{9.24}$$

uniformly for $1 \le i \le n$, where

$$\mathbf{P}_{\mathbf{a}}(U_i) = \Big[\dot{p}_{\lambda_{21}}(\|\widetilde{\boldsymbol{\alpha}}_1\|)\frac{\widehat{a}_1(U_i)}{\|\widehat{\boldsymbol{\alpha}}_1\|}, \cdots, \dot{p}_{\lambda_{2d_n(1)}}(\|\widetilde{\boldsymbol{\alpha}}_{d_n(1)}\|)\frac{\widehat{a}_{d_n(1)}(U_i)}{\|\widehat{\boldsymbol{\alpha}}_{d_n(1)}\|}\Big]^{\mathrm{T}}$$

and

$$\mathbf{P}_{\mathbf{b}}(U_i) = \Big[\dot{p}_{\lambda_{11}}(\|\widetilde{\boldsymbol{\beta}}_1\|)\frac{\widehat{d}_1(U_i)}{\|\widehat{\boldsymbol{\beta}}_1\|}, \cdots, \dot{p}_{\lambda_{1d_n(1)}}(\|\widetilde{\boldsymbol{\beta}}_{d_n(1)}\|)\frac{\widehat{d}_{d_n(1)}(U_i)}{\|\widehat{\boldsymbol{\beta}}_{d_n(1)}\|}\Big]^{\mathrm{T}},$$

$\widehat{d}_j(U_i)$ is the i-th element of $\mathbf{b}_j$.

By Assumption A7, we can prove that

$$\frac{\sqrt{nh}}{n}\big\|\mathbf{P}_{\mathbf{a}}(U_i)\big\| \leq \frac{a_{n2}\sqrt{d_n(1)h}}{\sqrt{n}} = o_P(1) \tag{9.25}$$

and

$$\frac{\sqrt{nh}}{n}\big\|\mathbf{P}_{\mathbf{b}}(U_i)\big\| \leq \frac{a_{n1}\sqrt{d_n(1)h}}{\sqrt{n}} = o_P(1). \tag{9.26}$$

Then, by (9.23)–(9.26), and following standard argument in Wang and Xia (2009), we can prove (4.7). The proof of (4.8) is analogous, and details are omitted here. □

Proof of Corollary 4.1. Based on Theorem 4.2, Remark 4.2 in Chapter 4.1, Theorem 2 in Cai *et al* (2000) and Theorem 1 in Zhang and Peng (2010), we can easily prove (4.9) and (4.10).

□

9.3 Proofs of some auxiliary results

Proof of (9.5). Let $V_{nu}(k,l)$ be the (k,l)-th element of $\frac{1}{n}\ddot{\mathcal{L}}_{nu}(\mathbf{a}(u),\mathbf{b}(u))$, and $V_u(k,l)$ be the (k,l)-th element of $\Lambda(u)\otimes V_2(u)$. Then, by the uniform consistency result for nonparametric kernel estimation (c.f.,

Theorem B in Mack and Silverman 1982),

$$\sup_{u\in[0,1]} \left|V_{nu}(k,l) - V_u(k,l)\right| = O_P(h^2 + \sqrt{\frac{\log n}{nh}}) = O_P(\sqrt{\frac{\log n}{nh}})$$

as $h \propto n^{-1/3}$ in Assumption A5.

Note that

$$\begin{aligned}\frac{1}{n}\ddot{\mathcal{L}}_{nu}(\mathbf{a}(u),\mathbf{b}(u)) =& \Lambda(u)\otimes V_2(u) + \frac{1}{n}\ddot{\mathcal{L}}_{nu}(\mathbf{a}(u),\mathbf{b}(u)) - \Lambda(u)\otimes V_2(u)\\ =:& \Lambda(u)\otimes V_2(u) + \widetilde{V}_{nu}, \end{aligned} \tag{9.27}$$

where $\widetilde{V}_{nu}$ is a $2d_n \times 2d_n$ matrix with the (k,l)-th element being $V_{nu}(k,l) - V_u(k,l)$.

Recall that $\lambda_s(u)$ is the smallest eigenvalue of $-\Lambda(u)\otimes V_2(u)$ and $\lambda_s = \inf_{u\in[0,1]} \lambda_s(u)$. By Assumption A2(iii), $\lambda_s > 0$. Thus, in order to prove (9.5), it suffices to show that the largest eigenvalue of $\widetilde{V}_{nu}$ is $o(1)$ in probability. Let $\widetilde{\lambda}_n(u)$ be the largest eigenvalue of $\widetilde{V}_{nu}$ and $\widetilde{\lambda}_n = \sup_{u\in[0,1]} \widetilde{\lambda}_n(u)$. Note that, by Assumption A5,

$$\widetilde{\lambda}_n(u) \le \max_k \sum_{l=1}^{2d_n} \left[V_{nu}(k,l) - V_u(k,l)\right] = O_P(d_n\sqrt{\frac{\log n}{nh}}) = o_P(1)$$

uniformly for $u \in [0,1]$. We then complete the proof of (9.5). □

Proof of (9.11). Recall that $\dot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n) = \mathbf{0}$ by the definition of the

local maximum likelihood estimation. Note that

$$
\begin{aligned}
I_{n1} =& \mathcal{L}_{n*}\big(\mathcal{A}_0 + \gamma_n \mathcal{U}, \mathcal{B}_0 + \gamma_n \mathcal{V}/h\big) - \mathcal{L}_{n*}\big(\mathcal{A}_0, \mathcal{B}_0\big) \\
=& \Big\{ \frac{1}{2} \Big[\begin{array}{c} \mathcal{A}_0 - \widetilde{\mathcal{A}}_n + \gamma_n \mathcal{U} \\ h\big(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n\big) + \gamma_n \mathcal{V} \end{array} \Big]^{\mathrm{T}} \ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n) \Big[\begin{array}{c} \mathcal{A}_0 - \widetilde{\mathcal{A}}_n + \gamma_n \mathcal{U} \\ h\big(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n\big) + \gamma_n \mathcal{V} \end{array} \Big] \\
& - \frac{1}{2} \Big[\begin{array}{c} \mathcal{A}_0 - \widetilde{\mathcal{A}}_n \\ h\big(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n\big) \end{array} \Big]^{\mathrm{T}} \ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n) \Big[\begin{array}{c} \mathcal{A}_0 - \widetilde{\mathcal{A}}_n \\ h\big(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n\big) \end{array} \Big] \Big\} \\
& + \Big\{ \big[\dot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n)\big]^{\mathrm{T}} \Big[\begin{array}{c} \mathcal{A} - \widetilde{\mathcal{A}}_n + \gamma_n \mathcal{U} \\ h\big(\mathcal{B} - \widetilde{\mathcal{B}}_n\big) + \gamma_n \mathcal{V} \end{array} \Big] - \big[\dot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n)\big]^{\mathrm{T}} \Big[\begin{array}{c} \mathcal{A}_0 - \widetilde{\mathcal{A}}_n \\ h\big(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n\big) \end{array} \Big] \Big\} \\
=:& I_{n1}(1) + I_{n1}(2).
\end{aligned}
$$

By Taylor's expansion, we have

$$
\begin{aligned}
I_{n1}(2) =& \gamma_n \big[\dot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n)\big]^{\mathrm{T}} \Big(\begin{array}{c} \mathcal{U} \\ \mathcal{V} \end{array} \Big) \overset{P}{\sim} \gamma_n \big[\dot{\mathcal{L}}_n(\mathcal{A}_0, \mathcal{B}_0)\big]^{\mathrm{T}} \Big(\begin{array}{c} \mathcal{U} \\ \mathcal{V} \end{array} \Big) \\
& - \gamma_n \Big[\begin{array}{c} \mathcal{A}_0 - \widetilde{\mathcal{A}}_n \\ h\big(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n\big) \end{array} \Big]^{\mathrm{T}} \big[\ddot{\mathcal{L}}_n(\mathcal{A}_0, \mathcal{B}_0)\big]^{\mathrm{T}} \Big(\begin{array}{c} \mathcal{U} \\ \mathcal{V} \end{array} \Big).
\end{aligned}
$$

On the other hand, by some elementary calculations, we also have

$$
\begin{aligned}
I_{n1}(1) =& \Big\{ \frac{1}{2} \Big[\begin{array}{c} \mathcal{A}_0 - \widetilde{\mathcal{A}}_n + \gamma_n \mathcal{U} \\ h\big(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n\big) + \gamma_n \mathcal{V} \end{array} \Big]^{\mathrm{T}} \ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n) \Big[\begin{array}{c} \mathcal{A}_0 - \widetilde{\mathcal{A}}_n + \gamma_n \mathcal{U} \\ h\big(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n\big) + \gamma_n \mathcal{V} \end{array} \Big] \\
& - \frac{1}{2} \Big[\begin{array}{c} \mathcal{A}_0 - \widetilde{\mathcal{A}}_n \\ h\big(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n\big) \end{array} \Big]^{\mathrm{T}} \ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n) \Big[\begin{array}{c} \mathcal{A}_0 - \widetilde{\mathcal{A}}_n + \gamma_n \mathcal{U} \\ h\big(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n\big) + \gamma_n \mathcal{V} \end{array} \Big] \Big\} \\
& + \Big\{ \frac{1}{2} \Big[\begin{array}{c} \mathcal{A}_0 - \widetilde{\mathcal{A}}_n \\ h\big(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n\big) \end{array} \Big]^{\mathrm{T}} \ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n) \Big[\begin{array}{c} \mathcal{A}_0 - \widetilde{\mathcal{A}}_n + \gamma_n \mathcal{U} \\ h\big(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n\big) + \gamma_n \mathcal{V} \end{array} \Big] \\
& - \frac{1}{2} \Big[\begin{array}{c} \mathcal{A}_0 - \widetilde{\mathcal{A}}_n \\ h\big(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n\big) \end{array} \Big]^{\mathrm{T}} \ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n) \Big[\begin{array}{c} \mathcal{A}_0 - \widetilde{\mathcal{A}}_n \\ h\big(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n\big) \end{array} \Big] \Big\} \\
=& \frac{\gamma_n}{2} \big[\mathcal{U}^{\mathrm{T}}, \ \mathcal{V}^{\mathrm{T}} \big] \ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n) \Big[\begin{array}{c} \mathcal{A}_0 - \widetilde{\mathcal{A}}_n + \gamma_n \mathcal{U} \\ h\big(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n\big) + \gamma_n \mathcal{V} \end{array} \Big] \\
& + \frac{\gamma_n}{2} \Big[\begin{array}{c} \mathcal{A}_0 - \widetilde{\mathcal{A}}_n \\ h\big(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n\big) \end{array} \Big]^{\mathrm{T}} \ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n) \Big[\begin{array}{c} \mathcal{U} \\ \mathcal{V} \end{array} \Big] \\
=& \frac{\gamma_n^2}{2} \big[\mathcal{U}^{\mathrm{T}}, \ \mathcal{V}^{\mathrm{T}} \big] \ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n) \Big[\begin{array}{c} \mathcal{U} \\ \mathcal{V} \end{array} \Big] + \gamma_n \Big[\begin{array}{c} \mathcal{A}_0 - \widetilde{\mathcal{A}}_n \\ h\big(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n\big) \end{array} \Big]^{\mathrm{T}} \ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n) \Big[\begin{array}{c} \mathcal{U} \\ \mathcal{V} \end{array} \Big] \\
\overset{P}{\sim}& \frac{\gamma_n^2}{2} \big[\mathcal{U}^{\mathrm{T}}, \ \mathcal{V}^{\mathrm{T}} \big] \ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n) \Big[\begin{array}{c} \mathcal{U} \\ \mathcal{V} \end{array} \Big] + \gamma_n \Big[\begin{array}{c} \mathcal{A}_0 - \widetilde{\mathcal{A}}_n \\ h\big(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n\big) \end{array} \Big]^{\mathrm{T}} \ddot{\mathcal{L}}_n \mathcal{A}_0, \mathcal{B}_0) \Big[\begin{array}{c} \mathcal{U} \\ \mathcal{V} \end{array} \Big].
\end{aligned}
$$

We can easily prove (9.11) by using the above results. □

Proof of (9.12). Recall that

$$
I_{n4} = \gamma_n \big(\mathcal{U}^{\mathrm{T}}, \mathcal{V}^{\mathrm{T}} \big) \dot{\mathcal{L}}_n(\mathcal{A}_0, \mathcal{B}_0). \tag{9.28}
$$

By Taylor's expansion for $q_1(r, y)$ and Assumption A4, we have

$$\begin{aligned} & q_1\Big\{ \sum_{j=1}^{d_n} \big[a_j(U_k) + \dot{a}_j(U_k)(U_i - U_k)\big] x_{ij},\ y_i \Big\} \\ =& q_1\Big[\sum_{j=1}^{d_n} a_j(U_i) x_{ij},\ y_i \Big] + O_P(h^2), \end{aligned} \tag{9.29}$$

which implies that

$$\begin{aligned} I_{n4} =& \gamma_n \sum_{k=1}^{n} \sum_{i=1}^{n} q_1\big[\sum_{j=1}^{d_n} a_j(U_i) x_{ij},\ y_i \big] X_i^{\mathrm{T}} \mathbf{u}(k) K_h(U_i - U_k) \\ & + \gamma_n \sum_{k=1}^{n} \sum_{i=1}^{n} q_1\Big[\sum_{j=1}^{d_n} a_j(U_i) x_{ij},\ y_i \Big] \Big(\frac{U_i - U_k}{h} \Big) X_i^{\mathrm{T}} \mathbf{v}(k) K_h(U_i - U_k) \\ & + O_P(\gamma_n n^{3/2} h^2) \cdot \big(\|\mathcal{U}\| + \|\mathcal{V}\| \big). \end{aligned} \tag{9.30}$$

Note that (U_i, X_i, y_i), $i = 1, \cdots, n$, are independent and identically distributed. By Assumptions A1–A3 and the Cauchy-Schwarz

inequality, we have

$$
\begin{aligned}
&\mathsf{E}\Big[\sum_{k=1}^{n}\sum_{i=1}^{n} q_1\big[\sum_{j=1}^{d_n} a_j(U_i)x_{ij},\ y_i\big]X_i^{\mathrm{T}}\mathbf{u}(k)K_h(U_i-U_k)\Big]^2 \qquad (9.31)\\
\le& n\sum_{k=1}^{n}\mathsf{E}\Big\{\sum_{i=1}^{n} q_1\big[\sum_{j=1}^{d_n} a_j(U_i)x_{ij},\ y_i\big]X_i^{\mathrm{T}}\mathbf{u}(k)K_h(U_i-U_k)\Big\}^2\\
=& n\sum_{k=1}^{n}\mathsf{E}\Big[\mathsf{E}\Big(\Big\{\sum_{i=1}^{n} q_1\big[\sum_{j=1}^{d_n} a_j(U_i)x_{ij},\ y_i\big]X_i^{\mathrm{T}}\mathbf{u}(k)K_h(U_i-U_k)\Big\}^2\Big|U_k\Big)\Big]\\
=& n\sum_{k=1}^{n}\sum_{i=1}^{n}\mathsf{E}\Big[\mathsf{E}\Big(\Big\{q_1^2\big[\sum_{j=1}^{d_n} a_j(U_i)x_{ij},\ y_i\big]\Big\}\mathbf{u}^{\mathrm{T}}(k)X_iX_i^{\mathrm{T}}\mathbf{u}(k)K_h^2(U_i-U_k)\Big|U_k\Big)\Big]\\
=& O\Big(n^2h^{-1}\sum_{k=1}^{n}\mathbf{u}^{\mathrm{T}}(k)\mathbf{u}(k)\Big) = O(n^2h^{-1})\cdot\|\mathcal{U}\|^2.
\end{aligned}
$$

Similarly, we can also show that

$$
\begin{aligned}
&\mathsf{E}\big[\sum_{k=1}^{n}\sum_{i=1}^{n} q_1\Big[\sum_{j=1}^{d_n} a_j(U_i)x_{ij},\ y_i\Big]\Big(\frac{U_i-U_k}{h}\Big)X_i^{\mathrm{T}}\mathbf{v}(k)K_h(U_i-U_k)\Big]^2\\
&= O(n^2h^{-1})\cdot\|\mathcal{V}\|^2.
\end{aligned}
$$

Thus, by (9.30) and noting $h \propto n^{-1/3}$ in Assumption A5, we have

$$
I_{n4} = O_P(\gamma_n^2 n^{3/2})\cdot\big(\|\mathcal{U}\|+\|\mathcal{V}\|\big). \qquad (9.32)
$$

We then complete the proof of (9.12). □

Verification of Assumption A6: We next show that Assumption A6 can be satisfied for LASSO and SCAD penalty functions with certain mild restrictions.

If the penalty function is the LASSO penalty defined by $p_\lambda(\cdot) =$

$n\lambda|\cdot|$, it is easy to see that Assumption A6(i) is satisfied. Note that

$$a_{n1} = n \max_{1\le j\le d_n(1)} \lambda_{1j}, \quad a_{n2} = n \max_{1\le j\le d_n(2)} \lambda_{2j}$$

and

$$b_{n1} = n \min_{d_n(1)+1\le j\le d_n} \lambda_{1j}, \quad b_{n2} = n \min_{d_n(2)+1\le j\le d_n} \lambda_{2j}.$$

By Assumption A5 and the definition of γ_n, we can show that Assumption A6(ii) is satisfied if

$$\max_{1\le j\le d_n(1)} \lambda_{1j} = o(n^{-1/10}) \text{ and } \max_{1\le j\le d_n(2)} \lambda_{2j} = o(n^{1/10}).$$

We can further show that Assumption A6(iii) is satisfied if

$$\frac{\sqrt{d_n}n^{1/10}}{\min_{d_n(1)+1\le j\le d_n} \lambda_{1j}} + \frac{\sqrt{d_n}n^{1/10}}{\min_{d_n(2)+1\le j\le d_n} \lambda_{2j}} = o(1). \tag{9.33}$$

We next consider the SCAD penalty function defined by $p_\lambda(\cdot) = n\rho_\lambda(\cdot)$, with

$$\dot\rho_\lambda(|z|) = \lambda I(|z|\le\lambda) + \frac{(a\lambda - |z|)_+}{a-1} I(|z|>\lambda), \tag{9.34}$$

where $a = 3.7$ as suggested by Fan and Li (2001). It is easy to check that Assumption A6(i) is satisfied. If we assume that

$$\max_{1\le j\le d_n(1)} \lambda_{1j} + \max_{1\le j\le d_n(2)} \lambda_{2j} = o(n^{1/2}),$$

by Proposition 4.1, we may show that $a_{n1} = a_{n2} = 0$ with probability 1, which indicates that Assumption A6(ii) is satisfied. By using the

definition of the SCAD penalty function, we can further show that Assumption A6(iii) is satisfied if (9.33) holds.

10 Proofs of the theoretical results in Chapter 4.2

In Chapter 10.1, we give some assumptions which are needed to prove the asymptotic theory in Chapter 4.2. In Chapter 10.2 and Chapter 10.3, we provide the proofs of the main theoretical results and some technical lemmas, respectively.

10.1 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

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

for $l=0,1,2$. For some sufficiently large $b_0>1$, let

$$\begin{aligned}\boldsymbol{\Omega}_0(b_0)=\Big\{&\mathbf{v}=(v_{11},\cdots,v_{1d_n},v_{21},\cdots,v_{2d_n})^{\mathrm{T}}:\ \|\mathbf{v}\|=1,\\ &\sum_{j=1}^{d_n}(|v_{1j}|+|v_{2j}|)\le b_0\sum_{j=1}^{s_{n2}}(|v_{1j}|+|v_{2j}|)\Big\}.\end{aligned}$$

Let $\alpha_n \propto \beta_n$ denote $b_1\beta_n \leq \alpha_n \leq b_2\beta_n$ when n is sufficiently large, $0 < b_1 \leq b_2 < \infty$, and let $\alpha_n \ll \beta_n$ denote $\alpha_n = o(\beta_n)$. We next introduce some assumptions which have been used in Chapter 4.2 to establish the asymptotic theory for the proposed feature selection and model specification procedure. Some of the conditions might be not the weakest possible conditions.

Assumption B1. The kernel function $K(\cdot)$ is a continuous and symmetric probability density function with a compact support.

Assumption B2. *(i)* Let

$$\mathsf{E}\Big\{q_1\big[\sum_{j=1}^{d_n} a_j(U_i)x_{ij}, y_i\big]\big|X_i, U_i\Big\} = 0 \quad a.s.,$$

and $\mathsf{E}\Big\{q_1^2\big[\sum_{j=1}^{d_n} a_j(U)x_j, y\big]|U = u\Big\}$ be continuous for $u \in [0,\ 1]$. Moreover, suppose that either

$$\max_{1\leq j\leq d_n} \mathsf{E}\Big\{\big|q_1\big[\sum_{j_1=1}^{d_n} a_{j_1}(U_i)x_{ij_1}, y_i\big]x_{ij}\big|^{m_0}\Big\} < \infty \tag{10.1}$$

for $m_0 > 2$, or

$$\max_{1\leq j\leq d_n} \mathsf{E}\Big\{\big|q_1\big[\sum_{j_1=1}^{d_n} a_{j_1}(U_i)x_{ij_1}, y_i\big]x_{ij}\big|^{m}\Big\} \leq \frac{M_0 m!}{2} \tag{10.2}$$

for all $m \geq 2$ 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 a $M(X, U, y) > 0$ such that

$$\Big|q_2\big[r_2(X,U),y\big]-q_2\big[r_1(X,U),y\big]\Big| \le M(X,U,y)\Big|r_2(X,U)-r_1(X,U)\Big|$$

and

$$\max_{i,j,k}\sup_{u\in[0,1]}\mathsf{E}\Big[\big|x_ix_jx_k\big|M(X,U,y)\Big|U=u\Big]<\infty.$$

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

$$\rho_1 \le \inf_{u\in[0,1]}\inf_{\mathbf{v}\in\mathbf{\Omega}_0(c_0)}\mathbf{v}^{\mathrm{T}}\big[-\ddot{\mathcal{L}}_n(u)\big]\mathbf{v} \le \sup_{u\in[0,1]}\sup_{\mathbf{v}\in\mathbf{\Omega}_0(c_0)}\mathbf{v}^{\mathrm{T}}\big[-\ddot{\mathcal{L}}_n(u)\big]\mathbf{v} \le \rho_2$$

with probability approaching one.

Assumption B3. 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 B4. The functional coefficients, $a_j(\cdot)$, have continuous second-order derivatives for $j = 1, \cdots, d_n$.

Assumption B5. Let $d_n \propto n^{\tau_1}$ and $\frac{nh}{(nd_n)^{2/m_0}\log h^{-1}} \to \infty$, where $0 \le \tau_1 < \infty$ and m_0 is defined in (10.1). Moreover, the bandwidth h and the tuning parameters λ_1 and λ_2 satisfy $h^2 \ll \big(\frac{\log h^{-1}}{nh}\big)^{1/2}$, $\big(\frac{\log h^{-1}}{nh}\big)^{1/2} = o(\lambda_1)$, $\lambda_1 \propto \lambda_2$ and $s_{n2}\lambda_1^2 h^{-2} = o(1)$.

Assumption B5′. 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 λ_1 and λ_2 satisfy $h^2 \ll \big(\frac{\log h^{-1}}{nh}\big)^{\tau_3/2}$, $\big(\frac{\log h^{-1}}{nh}\big)^{\tau_3/2} = o(\lambda_1)$, $\lambda_1 \propto \lambda_2$

and $s_{n2}\lambda_1^2 h^{-2} = o(1)$.

Assumption B6. Let $s_{n2}h^2 \propto (nh)^{-1/2}$, $\lambda_3 \sim \lambda_3^*$,

$$\lambda_3 = o\big(n^{\kappa/2}h^{-1/2}\big), \tag{10.3}$$
$$\lambda_3 \gg \lambda_1^{\kappa}(ns_{n2})^{\kappa/2}h^{-1/2}\big[(\log h^{-1})^{1/2} + s_{n2}^{1/2}(1+\lambda_1\sqrt{nh})\big].$$

Furthermore, assume that

$$\big(\min_{1\le j\le s_{n2}} \|\boldsymbol{\alpha}_{j0}\| + \min_{1\le j\le s_{n1}} D_j\big) \ge b_0 n^{1/2}, \quad b_0 > 0. \tag{10.4}$$

Assumption B6′. Let $s_{n2}h^2 \propto (nh)^{-1/2}$, $\lambda_4 \sim \lambda_4^*$,

$$\lambda_4 = o\big(s_{n2}^{1/2}n^{1/2}\lambda_1\big), \quad \lambda_4 \gg h^{-1/2}\big[(\log h^{-1})^{1/2} + s_{n2}^{1/2}(1+\lambda_1\sqrt{nh})\big] \tag{10.5}$$

and (10.4) hold.

Remark B.1. The above assumptions are mild and justifiable. Assumption B1 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 studies. The compact support restriction on the kernel function is not essential and can be removed at the cost of more tedious proofs. Assumption B2 imposes some smoothness and moment conditions on $q_1(\cdot,\cdot)$ and $q_2(\cdot,\cdot)$, some of which are commonly used in local maximum likelihood estimation (c.f., Cai *et al*, 2000, Li and Liang, 2008). Two moment conditions (10.1) and (10.2) on $q_1\big[\sum_{j_1=1}^{d_n} a_{j_1}(U_i)x_{ij_1}, y_i\big]x_{ij}$ are imposed in Assumption B2(i), and they are used to handle the polynomially diverging dimen-

sion of X (in Assumption B5) and the exponentially diverging dimension of X (in Assumption B5$'$), 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 (10.2), which may possibly limit the applicability of the model selection methodology. Assumption B2(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 B3 and B4 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 B5 imposes some restrictions on the bandwidth h and the tuning parameters λ_1 and λ_2 when $d_n \propto n^{\tau_1}$, whereas Assumption B5$'$ imposes some conditions when $d_n \propto \exp\left\{(nh)^{\tau_2}\right\}$. They are crucial to derive the uniform convergence rates for the preliminary estimation in Proposition 4.3. Noting that $h^2 \ll \left(\frac{\log h^{-1}}{nh}\right)^{1/2} = o(\lambda_1)$ and $\lambda_1 \propto \lambda_2$ by Assumption B5, the influence by h and λ_2 on the uniform convergence rate in (4.11) is dominated by that of λ_1. The Assumptions B6 and B6$'$ are mainly used to prove the sparsity and oracle property for the proposed feature selection and model specification procedure.

10.2 Proofs of the main results

We next give the detailed proofs of the main theoretical results developed in Chapter 4.2.

Proof of Proposition 4.3 (i). Recall that

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

The basic idea used in the proof of this proposition is similar to that in Bickel *et al* (2009) and Lian (2012). However, as we need to derive the uniform convergence rates for the kernel-based estimators, the technical argument would be more complicated. We start with the proof that with probability approaching one, uniformly for $k = 1, \cdots, n$,

$$\max\Big\{ \sum_{j=s_{n2}+1}^{d_n} |d_{jk}|, \sum_{j=s_{n1}+1}^{d_n} |\dot{d}_{jk}| \Big\} \le (1 + C_1)\Big(\sum_{j=1}^{s_{n2}} |d_{jk}| + \sum_{j=1}^{s_{n1}} |\dot{d}_{jk}| \Big), \tag{10.6}$$

where $C_1 > 0$ can be sufficiently large but independent of k, where $d_{jk} = \widetilde{a}_j(U_k) - a_j(U_k)$ and $\dot{d}_{jk} = h\big[\widetilde{\dot{a}}_j(U_k) - \dot{a}_j(U_k)\big]$, $j = 1, \cdots, d_n$, $k = 1, \cdots, n$.

By the definitions of $\widetilde{\mathbf{a}}_k$ and $\widetilde{\mathbf{b}}_k$, we readily have

$$\mathcal{Q}_{nk}(\widetilde{\mathbf{a}}_k, \widetilde{\mathbf{b}}_k) \ge \mathcal{Q}_{nk}\big(\mathbf{a}_{k0}, \mathbf{b}_{k0}\big), \tag{10.7}$$

where $\mathbf{a}_{k0}$ and $\mathbf{b}_{k0}$ are defined in Chapter 3.3. From (10.7), we have

$$\begin{aligned}&\mathcal{L}_{nk}(\widetilde{\mathbf{a}}_k, \widetilde{\mathbf{b}}_k) - \mathcal{L}_{nk}\big(\mathbf{a}_{k0}, \mathbf{b}_{k0}\big) \qquad (10.8)\\ \geq&\lambda_1\Big[\sum_{j=1}^{d_n} |\widetilde{a}_j(U_k)| - \sum_{j=1}^{d_n} |a_j(U_k)|\Big] + \lambda_2\Big[\sum_{j=1}^{d_n} |\widetilde{\dot{a}}_j(U_k)| - \sum_{j=1}^{d_n} |\dot{a}_j(U_k)|\Big].\end{aligned}$$

By the concavity condition of $\ell(\cdot,\cdot)$ (c.f., Assumption B2(ii)), we may show that

$$\mathcal{L}_{nk}(\widetilde{\mathbf{a}}_k, \widetilde{\mathbf{b}}_k) - \mathcal{L}_{nk}\big(\mathbf{a}_{k0}, \mathbf{b}_{k0}\big) \leq \mathbf{d}_k^{\mathrm{T}} \dot{\mathcal{L}}_{nk}, \qquad (10.9)$$

where

$$\begin{aligned}\dot{\mathcal{L}}_{nk} =&\frac{1}{n}\sum_{i=1}^{n} q_1\big[\sum_{j=1}^{d_n} a_j(U_k) + \dot{a}_j(U_k)(U_i - U_k)x_{ij}, y_i\big]\cdot\\ &\begin{pmatrix} X_i \\ \frac{U_i - U_k}{h}\cdot X_i \end{pmatrix} K_h(U_i - U_k)\end{aligned}$$

and $\mathbf{d}_k = (d_{1k}, \cdots, d_{d_nk}, \dot{d}_{1k}, \cdots, \dot{d}_{d_nk})^{\mathrm{T}}$. By Lemma 10.1 in Chapter 10.3, we may show that

$$\begin{aligned}&\max_{1\leq j\leq d_n}\sup_{1\leq k\leq n}\left|\frac{1}{n}\sum_{i=1}^{n} q_1\big[\sum_{j_1=1}^{d_n} a_{j_1}(U_i)x_{ij_1}, y_i\big]x_{ij}K_h(U_i - U_k)\right|\\ &\qquad = O_P\left(\sqrt{\frac{\log h^{-1}}{nh}}\right) \qquad (10.10)\end{aligned}$$

and

$$\max_{1\le j\le d_n}\sup_{1\le k\le n}\left|\frac{1}{n}\sum_{i=1}^{n}q_1\big[\sum_{j_1=1}^{d_n}a_{j_1}(U_i)x_{ij_1},y_i\big]x_{ij}\big(\frac{U_i-U_k}{h}\big)K_h(U_i-U_k)\right|$$
$$=O_P\left(\sqrt{\frac{\log h^{-1}}{nh}}\right). \tag{10.11}$$

Then, by (10.10), (10.11), the Cauchy-Schwarz inequality and the standard calculation in kernel-based smoothing, we may show that

$$\mathbf{d}_k^{\mathrm{T}}\dot{\mathcal{L}}_{nk}\le O_P\Big(\sqrt{\frac{\log h^{-1}}{nh}}+h^2\Big)\cdot\big(\sum_{j=1}^{d_n}|d_{jk}|+\sum_{j=1}^{d_n}|\dot{d}_{jk}|\big) \tag{10.12}$$

uniformly for $k=1,\cdots,n$.

On the other hand, by the triangle inequality, we may prove that

$$\begin{aligned}&\lambda_1\Big[\sum_{j=1}^{d_n}|\widetilde{a}_j(U_k)|-\sum_{j=1}^{d_n}|a_j(U_k)|\Big]\\ =&\lambda_1\sum_{j=1}^{s_{n2}}\big(|\widetilde{a}_j(U_k)|-|a_j(U_k)|\big)+\lambda_1\sum_{j=s_{n2}+1}^{d_n}|\widetilde{a}_j(U_k)|\\ \ge&-\lambda_1\sum_{j=1}^{s_{n2}}|d_{jk}|+\lambda_1\sum_{j=s_{n2}+1}^{d_n}|d_{jk}|.\end{aligned} \tag{10.13}$$

Similarly, we also have

$$\lambda_2\Big[\sum_{j=1}^{d_n}|\widetilde{\dot{a}}_j(U_k)|-\sum_{j=1}^{d_n}|\dot{a}_j(U_k)|\Big]\ge-\lambda_2\sum_{j=1}^{s_{n1}}|\dot{d}_{jk}|+\lambda_2\sum_{j=s_{n1}+1}^{d_n}|\dot{d}_{jk}|. \tag{10.14}$$

By (10.8), (10.9), (10.12)–(10.14) and the condition that $\sqrt{\frac{\log h^{-1}}{nh}}+$

$h^2 = o(\lambda_1 + \lambda_2)$ and $\lambda_1 \propto \lambda_2$, we can complete the proof of (10.6).

Let $\mathbf{u}_1$ and $\mathbf{u}_2$ be two d_n-dimensional column vectors and define

$$\mathbf{\Omega}(C_2) = \Big\{(\mathbf{u}_1^{\mathrm{T}},\ \mathbf{u}_2^{\mathrm{T}})^{\mathrm{T}} :\ \|\mathbf{u}_1\|^2 = \|\mathbf{u}_2\|^2 = C_2\Big\},$$

where C_2 is a positive constant. By the concavity of $\ell(\cdot,\cdot)$, we only need to prove that there exists a local maximiser $(\widetilde{\mathbf{a}}_k,\ h\widetilde{\mathbf{b}}_k)$ in the interior of the ball $\big\{(\mathbf{a}_{k0}+\gamma_n\mathbf{u}_1,\ h\mathbf{b}_{k0}+\gamma_n\mathbf{u}_2) :\ \ (\mathbf{u}_1^{\mathrm{T}},\ \mathbf{u}_2^{\mathrm{T}})^{\mathrm{T}} \in \mathbf{\Omega}(C_2)\big\}$, where $\gamma_n = \sqrt{s_{n2}}\lambda_1$. For simplicity, in the sequel, we let $\mathbf{u}_1 = \widetilde{\mathbf{a}}_k - \mathbf{a}_{k0}$ and $\mathbf{u}_2 = h(\widetilde{\mathbf{b}}_k - \mathbf{b}_{k0})$. Observe that

$$\mathcal{Q}_{nk}\big[\mathbf{a}_{k0}+\gamma_n\mathbf{u}_1, \mathbf{b}_{k0}+\gamma_n\mathbf{u}_2/h\big] - \mathcal{Q}_{nk}\big(\mathbf{a}_{k0}, \mathbf{b}_{k0}\big) = \sum_{l=1}^{3}\mathcal{I}_{nk}(l), \quad (10.15)$$

where

$$\begin{aligned}
\mathcal{I}_{nk}(1) =& \mathcal{L}_{nk}\big[\mathbf{a}_{k0}+\gamma_n\mathbf{u}_1, \mathbf{b}_{k0}+\gamma_n\mathbf{u}_2/h\big] - \mathcal{L}_{nk}\big(\mathbf{a}_{k0}, \mathbf{b}_{k0}\big),\\
\mathcal{I}_{nk}(2) =& -\lambda_1\Big[\sum_{j=1}^{d_n}|a_j(U_k)+\gamma_n u_{1j}| - \sum_{j=1}^{d_n}|a_j(U_k)|\Big],\\
\mathcal{I}_{nk}(3) =& -\lambda_2\Big[\sum_{j=1}^{d_n}|h\dot{a}_j(U_k)+\gamma_n u_{2j}| - \sum_{j=1}^{d_n}|h\dot{a}_j(U_k)|\Big],
\end{aligned}$$

in which u_{1j} and u_{2j} are the j-th element of $\mathbf{u}_1$ and $\mathbf{u}_2$, respectively.

We first consider $\mathcal{I}_{nk}(1)$. Letting $\mathbf{u} = (\mathbf{u}_1^{\mathrm{T}},\ \mathbf{u}_2^{\mathrm{T}})^{\mathrm{T}}$ and by the definition of $\mathcal{L}_{nk}(\cdot,\cdot)$ in Chapter 3.1, we have

$$\mathcal{I}_{nk}(1) \overset{P}{\sim} \gamma_n\mathbf{u}^{\mathrm{T}}\dot{\mathcal{L}}_{nk} + \frac{1}{2}\gamma_n^2\mathbf{u}^{\mathrm{T}}\ddot{\mathcal{L}}_{nk}(\mathbf{a}_k^*, \mathbf{b}_k^*)\mathbf{u}, \quad (10.16)$$

where $a_n \overset{P}{\sim} b_n$ denotes that $a_n = b_n(1 + o_P(1))$, $(\mathbf{a}_k^*, \mathbf{b}_k^*)$ lies between $\big(\mathbf{a}_{k0} + \gamma_n \mathbf{u}_1, \mathbf{b}_{k0} + \gamma_n \mathbf{u}_2/h\big)$ and $(\mathbf{a}_{k0}, \mathbf{b}_{k0})$,

$$\ddot{\mathcal{L}}_{nk}(\mathbf{a}_k, \mathbf{b}_k) = \begin{bmatrix} \ddot{\mathcal{L}}_{nk}(\mathbf{a}_k, \mathbf{b}_k, 0) & \ddot{\mathcal{L}}_{nk}(\mathbf{a}_k, \mathbf{b}_k, 1) \\ \ddot{\mathcal{L}}_{nk}(\mathbf{a}_k, \mathbf{b}_k, 1) & \ddot{\mathcal{L}}_{nk}(\mathbf{a}_k, \mathbf{b}_k, 2) \end{bmatrix}$$

with

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

for $l = 0, 1, 2$.

By (10.6), and noting that $\mathbf{u}_1 = (\widetilde{\mathbf{a}}_k - \mathbf{a}_{k0})/\gamma_n$ and $\mathbf{u}_2 = h(\widetilde{\mathbf{b}}_k - \mathbf{b}_{k0})/\gamma_n$, we may show that there exists $C_3 > 0$ such that

$$\sum_{j=1}^{d_n} (|u_{1j}| + |u_{2j}|) \le C_3 \sum_{j=1}^{s_{n2}} (|u_{1j}| + |u_{2j}|). \tag{10.17}$$

Using Lemma 10.1 in Chapter 10.3, the Cauchy-Schwarz inequality and (10.17), we can show that

$$\gamma_n \mathbf{u}^{\mathrm{T}} \dot{\mathcal{L}}_{nk} = O_P(\gamma_n^2) \cdot \|\mathbf{u}\|. \tag{10.18}$$

Note that

$$\begin{aligned}&\frac{1}{2}\gamma_n^2\mathbf{u}^{\mathrm{T}}\ddot{\mathcal{L}}_{nk}(\mathbf{a}_k^*,\mathbf{b}_k^*)\mathbf{u} \qquad (10.19)\\ =&\frac{1}{2}\gamma_n^2\mathbf{u}^{\mathrm{T}}\big[\ddot{\mathcal{L}}_{nk}(\mathbf{a}_k^*,\mathbf{b}_k^*)-\ddot{\mathcal{L}}_{nk}(\mathbf{a}_{k0},\mathbf{b}_{k0})\big]\mathbf{u}+\frac{1}{2}\gamma_n^2\mathbf{u}^{\mathrm{T}}\ddot{\mathcal{L}}_{nk}(\mathbf{a}_{k0},\mathbf{b}_{k0})\mathbf{u}.\end{aligned}$$

By Assumption B2(iii), we readily have

$$\frac{1}{2}\gamma_n^2\mathbf{u}^{\mathrm{T}}\ddot{\mathcal{L}}_{nk}(\mathbf{a}_{k0},\mathbf{b}_{k0})\mathbf{u}\le-\frac{1}{2}\rho_1\gamma_n^2\|\mathbf{u}\|^2<0. \qquad (10.20)$$

By Assumption B2(ii), we can prove that

$$\gamma_n^2\mathbf{u}^{\mathrm{T}}\big[\ddot{\mathcal{L}}_{nk}(\mathbf{a}_k^*,\mathbf{b}_k^*)-\ddot{\mathcal{L}}_{nk}(\mathbf{a}_{k0},\mathbf{b}_{k0})\big]\mathbf{u}=o_P(\gamma_n^2)\cdot\big(\|\mathbf{u}\|^2\big). \qquad (10.21)$$

Hence, by (10.16) and (10.18)–(10.21), when n is sufficiently large, by taking C_2 large enough, we have

$$\mathcal{I}_{nk}(1)\overset{P}{\sim}\frac{1}{2}\gamma_n^2\mathbf{u}^{\mathrm{T}}\ddot{\mathcal{L}}_{nk}(\mathbf{a}_{k0},\mathbf{b}_{k0})\mathbf{u}. \qquad (10.22)$$

We next consider $\mathcal{I}_{nk}(2)$ and $\mathcal{I}_{nk}(3)$. It is easy to show that

$$\begin{aligned}\mathcal{I}_{nk}(2)=&-\lambda_1\Big[\sum_{j=1}^{d_n}|a_j(U_k)+\gamma_nu_{1j}|-\sum_{j=1}^{d_n}|a_j(U_k)|\Big]\\ \le&\lambda_1\sum_{j=1}^{s_{n2}}\big[|a_j(U_k)|-|a_j(U_k)+\gamma_nu_{1j}|\big]-\lambda_1\sum_{j=s_{n2}+1}^{d_n}|\gamma_nu_{1j}|\\ =&O_P(\gamma_n^2)\cdot\|\mathbf{u}_1\|-\lambda_1\sum_{j=s_{n2}+1}^{d_n}|\gamma_nu_{1j}|. \qquad (10.23)\end{aligned}$$

Similarly, noting that $\lambda_1 \propto \lambda_2$ we also have

$$\mathcal{I}_{nk}(3) = O_P(\gamma_n^2) \cdot \|\mathbf{u}_2\| - \lambda_2 \sum_{j=s_{n1}+1}^{d_n} |\gamma_n u_{2j}|. \tag{10.24}$$

Hence, by (10.15) and (10.22)–(10.24), we can prove that the leading term of $\mathcal{I}_{nk}(1)+\mathcal{I}_{nk}(2)+\mathcal{I}_{nk}(3)$ is negative in probability (uniformly in k), which indicates that $(\widetilde{\mathbf{a}}_k,\ \widetilde{\mathbf{b}}_k)$ is indeed in the interior of ball defined previously for sufficiently large C_2, and thus completes the proof of Proposition 4.3 (i). □

Proof of Proposition 4.3 (ii). The proof is similar to that in the proof of Proposition 4.3 (i) with the role of Lemma 10.1 replaced by Lemma 10.2. □

Proof of Theorem 4.3 (i). We start with the proof of the convergence rates for the biased oracle estimators $\widehat{\mathcal{A}}_n^o$ and $\widehat{\mathcal{B}}_n^o$. According to the definition, we have

$$\left(\widehat{\mathcal{A}}_n^o,\ \widehat{\mathcal{B}}_n^o\right) = \arg\max \mathcal{Q}_n^1(\mathcal{A}^o,\ \mathcal{B}^o), \tag{10.25}$$

where $\mathcal{A}^o$ and $\mathcal{B}^o$ are defined as in Chapter 4.2. Let $\mathcal{A}_0$ and $\mathcal{B}_0$ be the vectors of the true functional coefficients and their derivative functions, and denote

$$\mathcal{U}_1 = \left[\mathbf{u}_1^{\mathrm{T}}(1), \cdots, \mathbf{u}_1^{\mathrm{T}}(n)\right]^{\mathrm{T}}, \quad \mathcal{U}_2 = \left[\mathbf{u}_2^{\mathrm{T}}(1), \cdots, \mathbf{u}_2^{\mathrm{T}}(n)\right]^{\mathrm{T}},$$

where both $\mathbf{u}_1(k)$ and $\mathbf{u}_2(k)$ are d_n-dimensional column vectors, $k = 1, \cdots, n$, the last $d_n - s_{n2}$ elements of $\mathbf{u}_1(k)$ and the last $d_n - s_{n1}$

elements of $\mathbf{u}_2(k)$ are zeroes. Define

$$\mathbf{\Omega}_n^*(C_4) = \big\{(\mathcal{U}_1^{\mathrm{T}},\ \mathcal{U}_2^{\mathrm{T}})^{\mathrm{T}} :\ \|\mathcal{U}_1\|^2 = \|\mathcal{U}_2\|^2 = nC_4\big\},$$

where C_4 is a positive constant which can be sufficiently large.

For $(\mathcal{U}_1^{\mathrm{T}},\ \mathcal{U}_2^{\mathrm{T}})^{\mathrm{T}} \in \mathbf{\Omega}_n^*(C_4)$, observe that

$$\begin{aligned}&\mathcal{Q}_n^1\big(\mathcal{A}_0 + \gamma_n^*\mathcal{U}_1, \mathcal{B}_0 + \gamma_n^*\mathcal{U}_2/h\big) - \mathcal{Q}_n^1(\mathcal{A}_0, \mathcal{B}_0)\\ &= \mathcal{I}_n(1) + \mathcal{I}_n(2) + \mathcal{I}_n(3),\end{aligned} \tag{10.26}$$

where $\gamma_n^* = \sqrt{s_{n2}/nh}$,

$$\begin{aligned}\mathcal{I}_n(1) =& \mathcal{L}_n^{\diamond}\big(\mathcal{A}_0 + \gamma_n^*\mathcal{U}_1, \mathcal{B}_0 + \gamma_n^*\mathcal{U}_2/h\big) - \mathcal{L}_n^{\diamond}(\mathcal{A}_0, \mathcal{B}_0),\\ \mathcal{I}_n(2) =& \lambda_3\sum_{j=1}^{d_n}\|\widetilde{\boldsymbol{\alpha}}_j\|^{-\kappa}\|\boldsymbol{\alpha}_{j0}\| - \lambda_3\sum_{j=1}^{d_n}\|\widetilde{\boldsymbol{\alpha}}_j\|^{-\kappa}\|\boldsymbol{\alpha}_{j0} + \gamma_n^*\mathbf{u}_{1j}\|,\\ \mathcal{I}_n(3) =& \lambda_3^*\sum_{j=1}^{d_n}|\widetilde{D}_j|^{-\kappa}\|h\boldsymbol{\beta}_{j0}\| - \lambda_3^*\sum_{j=1}^{d_n}|\widetilde{D}_j|^{-\kappa}\|h\boldsymbol{\beta}_{j0} + \gamma_n^*\mathbf{u}_{2j}\|,\end{aligned}$$

in which $\boldsymbol{\alpha}_{j0} = \big[a_j(U_1), \cdots, a_j(U_n)\big]^{\mathrm{T}}$, $\boldsymbol{\beta}_{j0} = \big[\dot{a}_j(U_1), \cdots, \dot{a}_j(U_n)\big]^{\mathrm{T}}$, $\mathbf{u}_{1j} = \big[u_{1j}(1), \cdots, u_{1j}(n)\big]^{\mathrm{T}}$, $\mathbf{u}_{2j} = \big[u_{2j}(1), \cdots, u_{2j}(n)\big]^{\mathrm{T}}$, $u_{1j}(k)$ and $u_{2j}(k)$ are the j-th component of vectors $\mathbf{u}_1(k)$ and $\mathbf{u}_2(k)$, respectively.

For $\mathcal{I}_n(1)$, by the definition of $\mathcal{L}_n^{\diamond}(\cdot,\cdot)$ in Chapter 3.3, we have

$$\mathcal{I}_n(1) \overset{P}{\sim} \gamma_n^*\mathcal{V}_n^{\mathrm{T}}(\mathcal{U}_1,\mathcal{U}_2)\dot{\mathcal{L}}_n(\mathcal{A}_0,\mathcal{B}_0) + \frac{1}{2}(\gamma_n^*)^2\mathcal{V}_n^{\mathrm{T}}(\mathcal{U}_1,\mathcal{U}_2)\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\widetilde{\mathcal{B}}_n)\mathcal{V}_n(\mathcal{U}_1,\mathcal{U}_2). \tag{10.27}$$

The detailed proof of (10.27) will be provided in Chapter 10.3 below.

We define

$$\begin{aligned}\mathcal{I}_n(4) =&\gamma_n^* \mathcal{V}_n^{\mathrm{T}}(\mathcal{U}_1, \mathcal{U}_2)\dot{\mathcal{L}}_n(\mathcal{A}_0, \mathcal{B}_0),\\ \mathcal{I}_n(5) =&\frac{1}{2}(\gamma_n^*)^2 \mathcal{V}_n^{\mathrm{T}}(\mathcal{U}_1, \mathcal{U}_2)\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n)\mathcal{V}_n(\mathcal{U}_1, \mathcal{U}_2).\end{aligned}$$

By some elementary but tedious calculations, we can show that

$$\mathcal{I}_n(4) = O_P\big((\gamma_n^*)^2 n^{1/2}\big) \cdot \big(\|\mathcal{U}\| + \|\mathcal{V}\|\big). \tag{10.28}$$

The detailed proof of (10.28) will be also given in Chapter 10.3 below. For $\mathcal{I}_n(5)$, note that

$$\begin{aligned}\mathcal{I}_n(5) =&\frac{1}{2}(\gamma_n^*)^2 \mathcal{V}_n^{\mathrm{T}}(\mathcal{U}_1, \mathcal{U}_2)\Big[\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n, \widetilde{\mathcal{B}}_n) - \ddot{\mathcal{L}}_n(\mathcal{A}_0, \mathcal{B}_0)\Big]\mathcal{V}_n(\mathcal{U}_1, \mathcal{U}_2)+\\ &\frac{1}{2}(\gamma_n^*)^2 \mathcal{V}_n^{\mathrm{T}}(\mathcal{U}_1, \mathcal{U}_2)\ddot{\mathcal{L}}_n(\mathcal{A}_0, \mathcal{B}_0)\mathcal{V}_n(\mathcal{U}_1, \mathcal{U}_2)\\ \equiv&\mathcal{I}_n(6) + \mathcal{I}_n(7).\end{aligned} \tag{10.29}$$

By Assumption B2(iii) and the definitions of $\mathcal{U}_1$ and $\mathcal{U}_2$, we may show that

$$\mathcal{I}_n(7) \le -\frac{1}{2}\rho_1(\gamma_n^*)^2\big(\|\mathcal{U}_1\|^2 + \|\mathcal{U}_2\|^2\big) < 0. \tag{10.30}$$

By Assumption B2(ii) and using Proposition 4.3, we can prove that

$$\mathcal{I}_n(6) = o_P\big((\gamma_n^*)^2\big) \cdot \big(\|\mathcal{U}_1\|^2 + \|\mathcal{U}_2\|^2\big), \tag{10.31}$$

which, together with (10.27)–(10.30), implies that $\mathcal{I}_n(7)$ is the leading term of $\mathcal{I}_n(1)$. Hence, when n is sufficiently large, by taking C_4 large

enough, we have

$$\mathcal{I}_n(1) \overset{P}{\sim} \frac{1}{2}(\gamma_n^*)^2 \mathcal{V}_n^{\mathrm{T}}(\mathcal{U}_1, \mathcal{U}_2) \ddot{\mathcal{L}}_n(\mathcal{A}_0, \mathcal{B}_0) \mathcal{V}_n(\mathcal{U}_1, \mathcal{U}_2). \tag{10.32}$$

We next consider $\mathcal{I}_n(2)$. By Proposition 4.3 and noting that $\mathbf{u}_{1j} = \mathbf{0}$ for $j = s_{n2+1}, \cdots, d_n$ and $\lambda_3 = o\big(n^{\kappa/2} h^{-1/2}\big)$ in (10.3), we have

$$\begin{aligned}
\mathcal{I}_n(2) =& \lambda_3 \sum_{j=1}^{d_n} \|\widetilde{\boldsymbol{\alpha}}_j\|^{-\kappa} \|\boldsymbol{\alpha}_{j0}\| - \lambda_3 \sum_{j=1}^{d_n} \|\widetilde{\boldsymbol{\alpha}}_j\|^{-\kappa} \|\boldsymbol{\alpha}_{j0} + \gamma_n^* \mathbf{u}_{1j}\| \\
=& \lambda_3 \sum_{j=1}^{d_n} \|\widetilde{\boldsymbol{\alpha}}_j\|^{-\kappa} \big(\|\boldsymbol{\alpha}_{j0}\| - \|\boldsymbol{\alpha}_{j0} + \gamma_n^* \mathbf{u}_{1j}\|\big) \\
=& \lambda_3 \sum_{j=1}^{s_{n2}} \|\widetilde{\boldsymbol{\alpha}}_j\|^{-\kappa} \big(\|\boldsymbol{\alpha}_{j0}\| - \|\boldsymbol{\alpha}_{j0} + \gamma_n^* \mathbf{u}_{1j}\|\big) \\
=& O_P\big(\lambda_3 n^{-\kappa/2} s_{n2}^{1/2} \gamma_n^*\big) \cdot \|\mathcal{U}_1\| = o_P\big((\gamma_n^*)^2\big) \cdot \|\mathcal{U}_1\|^2.
\end{aligned} \tag{10.33}$$

Similarly, we may also show that

$$\mathcal{I}_n(3) = O_P\big(\lambda_3^* n^{-\kappa/2} s_{n2}^{1/2} \gamma_n^*\big) \cdot \|\mathcal{U}_2\| = o_P\big((\gamma_n^*)^2\big) \cdot \|\mathcal{U}_2\|^2. \tag{10.34}$$

Hence, by (10.26) and (10.32)–(10.34), we can prove that the leading term of $\mathcal{I}_n(1) + \mathcal{I}_n(2) + \mathcal{I}_n(3)$ is negative in probability, which indicates that for any $\epsilon > 0$, there exists a sufficiently large $C_4 > 0$ such that

$$\mathsf{P}\left\{ \sup_{(\mathcal{U}_1, \mathcal{U}_2) \in \boldsymbol{\Omega}_n^*(C_4)} \mathcal{Q}_n^1\big(\mathcal{A}_0 + \gamma_n^* \mathcal{U}_1, \mathcal{B}_0 + \gamma_n^* \mathcal{U}_2 / h\big) < \mathcal{Q}_n^1(\mathcal{A}_0, \mathcal{B}_0) \right\} \geq 1 - \epsilon \tag{10.35}$$

for large n. Therefore, we may show that

$$\frac{1}{n}\big\|\widehat{\mathcal{A}}_n^o - \mathcal{A}_0\big\|^2 = \frac{s_{n2}}{nh}, \quad \frac{1}{n}\big\|\widehat{\mathcal{B}}_n^o - \mathcal{B}_0\big\|^2 = \frac{s_{n2}}{nh^3}. \tag{10.36}$$

which is (4.12) in Theorem 4.3 (i).

In order to complete the proof of Theorem 4.3 (i), we need to apply Lemma 10.3 which is given in Chapter 10.3. By the definition of the biased oracle estimators $\widehat{\mathcal{A}}_n^o$ and $\widehat{\mathcal{B}}_n^o$, it is easy to verify (10.57) and (10.58). We next only show the proof of (10.59) as the proof of (10.60) is similar. Under the moment condition (10.1) and $d_n \propto n^{\tau_1}$, we may show that when $\mathcal{A} = \widehat{\mathcal{A}}_n^o$ and $\mathcal{B} = \widehat{\mathcal{B}}_n^o$, the left hand side of (10.59) satisfies

$$\max_{s_{n2}+1\le j\le d_n} \|\dot{\mathcal{L}}_n^\diamond(\mathcal{A},\, \mathcal{B} \,|\, \boldsymbol{\alpha}_j)\| = O_P\big((h^{-1}\log h^{-1})^{1/2} + (s_{n2}h^{-1} + s_{n2}n\lambda_1^2)^{1/2}\big) \tag{10.37}$$

with $\dot{\mathcal{L}}_n^\diamond(\mathcal{A},\, \mathcal{B} \,|\, \boldsymbol{\alpha}_j)$ being the gradient vector of $\mathcal{L}_n^\diamond(\mathcal{A},\, \mathcal{B})$ with respect to $\boldsymbol{\alpha}_j$, whereas the right hand side of (10.59) satisfies

$$\begin{aligned} &\lambda_3 \min_{s_{n2}+1\le j\le d_n} \|\widetilde{\boldsymbol{\alpha}}_j\|^{-\kappa} = \lambda_3 \min_{s_{n2}+1\le j\le d_n} \|\widetilde{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_{j0}\|^{-\kappa} \\ =&\lambda_3\big[\max_{s_{n2}+1\le j\le d_n} \|\widetilde{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_{j0}\|\big]^{-\kappa} \\ \ge& c_\diamond\big(\lambda_3\lambda_1^{-\kappa}(ns_{n2})^{-\kappa/2}\big) \end{aligned} \tag{10.38}$$

by Proposition 4.3, where $c_\diamond > 0$. Using (10.37), (10.38) and Assumption B6, we may prove (10.59). Similarly, under the moment condition (10.2) and $d_n \propto \exp\{(nh)^{\tau_2}\}$, we can also prove (10.59).

Then, the proof of Theorem 4.3 (i) is completed by using Lemma

10.3. □

Proof of Theorem 4.3 (ii). By using Proposition 4.3, the definition of the SCAD function and Lemma 10.4, the proof is similar to the proof of Theorem 4.3 (i). Hence details are omitted here to save space. □

Proof of Theorem 4.4. The proof is similar to the proof of Theorem 2 in Wang and Xia (2009) with some modifications. Recall that $\overline{a}_j^o(U_k)$, $j = 1, \cdots, s_{n2}$, $k = 1, \cdots, n$, are the biased oracle estimators of $a_j(U_k)$, i.e., the maximisation of the objective function $\mathcal{Q}_n^2(\mathcal{A}^o,\ \mathcal{B}^o)$ with respect to $\mathcal{A}^o$, and define

$$\overline{c}_j^o = \frac{1}{n}\sum_{k=1}^{n} \overline{a}_j^o(U_k), \quad j = s_{n1}+1, \cdots, s_{n2}.$$

Let

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

and

$$\overline{\mathbf{C}}_n^o = \left(\overline{c}_{s_{n1}+1}^o, \cdots, \overline{c}_{s_{n2}}^o\right)^{\mathrm{T}}.$$

By Theorem 4.3, in order to prove (4.12) and (4.13), we only need to show that

$$\sqrt{nh}\mathbf{B}_n^{\mathrm{T}}\overline{\mathbf{D}}_n^o = o_P(1), \quad \sqrt{n}\mathbf{A}_n^{\mathrm{T}}\left(\overline{\mathbf{C}}_n^o - \mathbf{C}_n^{uo}\right) = o_P(1). \tag{10.39}$$

For $k = 1, \cdots, n$, denote

$$\mathbf{a}^{uo}(U_k) = \big[a_1^{uo}(U_k), \cdots, a_{s_{n2}}^{uo}(U_k), 0, \cdots, 0\big]^{\mathrm{T}},$$
$$\overline{\mathbf{a}}^{o}(U_k) = \big[\overline{a}_1^{o}(U_k), \cdots, \overline{a}_{s_{n2}}^{o}(U_k), 0, \cdots, 0\big]^{\mathrm{T}},$$

where the last $d_n - s_{n2}$ elements in the above two vectors are zeros, and let $\mathbf{b}^{uo}(U_k)$ and $\overline{\mathbf{b}}^{o}(U_k)$ be defined analogously. Then, using the first-order condition, we may show that the oracle estimates satisfy the following equation:

$$\mathbf{0} = \mathcal{R}_{s_{n2}}\dot{\mathcal{L}}_{nk}^{*}\big(\widetilde{\mathbf{a}}_k, \widetilde{\mathbf{b}}_k\big) + \mathcal{R}_{s_{n2}}\ddot{\mathcal{L}}_{nk}^{*}\big(\widetilde{\mathbf{a}}_k, \widetilde{\mathbf{b}}_k\big) \begin{bmatrix} \mathbf{a}^{uo}(U_k) - \widetilde{\mathbf{a}}_k \\ \mathbf{b}^{uo}(U_k) - \widetilde{\mathbf{b}}_k \end{bmatrix} \tag{10.40}$$

uniformly for $1 \le k \le n$, where $\mathcal{R}_{s_{n2}} = \big[I_{s_{n2}},\ N_{s_{n2}\times(2d_n - s_{n2})}\big]$ with I_s being an $s \times s$ identity matrix and $N_{r\times s}$ being a $r \times s$ null matrix.

Following the proof of Theorem 4.3, we can also show that the biased oracle estimates satisfy the following equation:

$$\mathbf{0} = \mathcal{R}_{s_{n2}}\dot{\mathcal{L}}_{nk}^{*}\big(\widetilde{\mathbf{a}}_k, \widetilde{\mathbf{b}}_k\big) + \mathcal{R}_{s_{n2}}\ddot{\mathcal{L}}_{nk}^{*}\big(\widetilde{\mathbf{a}}_k, \widetilde{\mathbf{b}}_k\big) \begin{bmatrix} \overline{\mathbf{a}}^{o}(U_k) - \widetilde{\mathbf{a}}_k \\ \overline{\mathbf{b}}^{o}(U_k) - \widetilde{\mathbf{b}}_k \end{bmatrix} - \mathcal{P}^{*}(U_k) \tag{10.41}$$

uniformly for $1 \le i \le n$, where

$$\mathcal{P}^{*}(U_k) = \Big(\dot{p}_{\lambda_4}(\|\widetilde{\boldsymbol{\alpha}}_1\|)\frac{\overline{a}_1^{o}(U_k)}{\|\overline{\boldsymbol{\alpha}}_1^{o}\|}, \cdots, \dot{p}_{\lambda_4}(\|\widetilde{\boldsymbol{\alpha}}_{s_{n2}}\|)\frac{\overline{a}_{s_{n2}}(U_k)}{\|\overline{\boldsymbol{\alpha}}_{s_{n2}}^{o}\|}\Big)^{\mathrm{T}},$$

$\overline{\boldsymbol{\alpha}}_j^{o} = \big[\overline{a}_j^{o}(U_1), \cdots, \overline{a}_j^{o}(U_n)\big]^{\mathrm{T}}$. By Proposition 4.3 and Assumption B6′,

we may show that

$$\min_{1\le j\le s_{n2}} \|\widetilde{\boldsymbol{\alpha}}_j\| \ge \min_{1\le j\le s_{n2}} \|\boldsymbol{\alpha}_{j0}\| - \max_{1\le j\le s_{n2}} \|\widetilde{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_{j0}\| \ge \frac{1}{2} b_0 \sqrt{n}$$

with probability approaching one, which together with (10.5), indicates that the penalty term $\mathcal{P}^*(U_k)$ in (10.41) is asymptotically negligible. Hence, by (10.40) and (10.41), we can complete the proof of (10.39).□

10.3 Proofs of some technical lemmas

Define

$$Z_{ij}(u,l) = q_1\big[\sum_{j_1=1}^{d_n} a_{j_1}(U_i)x_{ij_1}, y_i\big]x_{ij}\big(\frac{U_i-u}{h}\big)^l K_h(U_i-u), \quad u\in[0,1] \tag{10.42}$$

for $i = 1,\ldots,n$, $j = 1,\ldots,d_n$, $l = 0,1,2,\cdots$. Under different moment conditions on the random element $q_1\big[\sum_{j=1}^{d_n} a_j(U_i)x_{ij}, y_i\big]x_{ij}$, in Lemmas 10.1 and 10.2 below, we give the uniform consistency results of the nonparametric kernel-based estimators in the ultra-high dimensional case, which are of independent interest.

Lemma 10.1. Suppose that Assumptions B1 and B3 in Chapter 10.1 are satisfied. Moreover, suppose that the dimension $d_n \propto n^{\tau_1}$ with $0 \le \tau_1 < \infty$, $\mathsf{E}\Big\{q_1\big[\sum_{j=1}^{d_n} a_j(U_i)x_{ij}, y_i\big]\big|X_i, U_i\Big\} = 0$ a.s., the moment condition (10.1) holds for some $m_0 > 2$, and

$$h \propto n^{-\delta_1} \text{ with } 0<\delta_1<1, \quad \frac{nh}{(nd_n)^{2/m_0}\log h^{-1}} \to \infty. \tag{10.43}$$

Then we have, as $n \to \infty$,

$$\max_{1\le j\le d_n} \sup_{u\in[0,1]} \Big|\frac{1}{n}\sum_{i=1}^{n} Z_{ij}(u,l)\Big| = O_P\Big(\big(\frac{\log h^{-1}}{nh}\big)^{1/2}\Big) \tag{10.44}$$

for any $l = 0, 1, 2, \cdots$.

Proof of Lemma 10.1. For simplicity, let $\xi_n = \big(\frac{\log h^{-1}}{nh}\big)^{1/2}$. The main idea of proving (10.44) is to consider covering the interval [0, 1] by a finite number of subsets $U(k)$ which are centered at u_k with radius $r_n = \xi_n h^2$. Letting $\mathcal{N}_n$ be the total number of such subsets $U(k)$, $\mathcal{N}_n = O(r_n^{-1})$. It is easy to show that

$$\begin{aligned}
&\max_{1\le j\le d_n} \sup_{u\in[0,1]} \Big|\frac{1}{n}\sum_{i=1}^{n} Z_{ij}(u,l)\Big| \\
\le &\max_{1\le j\le d_n} \max_{1\le k\le \mathcal{N}_n} \Big|\frac{1}{n}\sum_{i=1}^{n} Z_{ij}(u_k,l)\Big| + \\
&\max_{1\le j\le d_n} \max_{1\le k\le \mathcal{N}_n} \sup_{u\in U(k)} \Big|\frac{1}{n}\sum_{i=1}^{n} Z_{ij}(u,l) - \frac{1}{n}\sum_{i=1}^{n} Z_{ij}(u_k,l)\Big| \\
\equiv &\Pi_{n1} + \Pi_{n2}.
\end{aligned} \tag{10.45}$$

By the continuity condition on $K(\cdot)$ in Assumption B1 and using the definition of r_n, we readily have

$$\Pi_{n2} = O_P\big(\frac{r_n}{h^2}\big) = O_P(\xi_n). \tag{10.46}$$

For Π_{n1}, we apply the truncation technique and the Bernstein inequality for i.i.d. random variables (c.f., Lemma 2.2.9 in van der Vaart

and Wellner, 1996) to obtain the convergence rate. Let

$$\begin{aligned} M_n =& M_1(nd_n)^{1/m_0}, \\ \overline{Z}_{ij}(u,l) =& Z_{ij}(u,l) I\Big\{ \big|q_1\big[\sum_{j=1}^{d_n} a_j(U_i)x_{ij}, y_i\big]x_{ij}\big| \le M_n \Big\}, \\ \text{and } \widetilde{Z}_{ij}(u,l) =& Z_{ij}(u,l) - \overline{Z}_{ij}(u,l), \end{aligned}$$

where $I\{\cdot\}$ is an indicator function. Hence we have

$$\begin{aligned} \Pi_{n1} \le & \max_{1\le j\le d_n} \max_{1\le k\le \mathcal{N}_n} \Big| \frac{1}{n}\sum_{i=1}^{n} \big\{ \overline{Z}_{ij}(u_k,l) - \mathsf{E}[\overline{Z}_{ij}(u_k,l)] \big\} \Big| + \\ & \max_{1\le j\le d_n} \max_{1\le k\le \mathcal{N}_n} \Big| \frac{1}{n}\sum_{i=1}^{n} \big\{ \widetilde{Z}_{ij}(u_k,l) - \mathsf{E}[\widetilde{Z}_{ij}(u_k,l)] \big\} \Big| \\ \equiv & \Pi_{n3} + \Pi_{n4}. \end{aligned} \tag{10.47}$$

Note that for $M_2 > 0$ and any $\varepsilon > 0$, by (10.43) and the Markov inequality,

$$\begin{aligned} \mathsf{P}\Big(\Pi_{n4} > M_2\xi_n\Big) \le & \mathsf{P}\left(\max_{1\le k\le \mathcal{N}_n} \max_{1\le i\le n, 1\le j\le d_n} \big|\widetilde{Z}_{ij}(u_k,l)\big| > 0 \right) \\ \le & \sum_{j=1}^{d_n}\sum_{i=1}^{n} \mathsf{P}\left(\big|q_1\big[\sum_{j_1=1}^{d_n} a_{j_1}(U_i)x_{ij_1}, y_i\big]x_{ij}\big| > M_n \right) \\ \le & M_1^{-m_0}\mathsf{E}\big[\big|q_1\big[\sum_{j_1=1}^{d_n} a_{j_1}(U_i)x_{ij_1}, y_i\big]x_{ij}\big|^{m_0}\big] < \varepsilon, \end{aligned}$$

if we choose $M_1 > \mathsf{E}\big[\big|q_1\big[\sum_{j_1=1}^{d_n} a_{j_1}(U_i)x_{ij_1}, y_i\big]x_{ij}\big|^{m_0}\big]^{1/m_0}\varepsilon^{-1/m_0}$. Then,

by letting ε be arbitrarily small, we can show that

$$\Pi_{n4} = O_P(\xi_n). \tag{10.48}$$

Note that

$$\left|\overline{Z}_{ij}(u_k, l) - \mathsf{E}[\overline{Z}_{ij}(u_k, l)]\right| \leq \frac{CM_n}{h} \tag{10.49}$$

and

$$\mathrm{Var}\left[\overline{Z}_{ij}(u_k, l)\right] \leq \frac{C}{h} \tag{10.50}$$

for some $C > 0$. By (10.44), (10.48), (10.49) and Lemma 2.2.9 in van der Vaart and Wellner (1996), we have

$$\begin{aligned}\mathsf{P}(\Pi_{n3} > M_2\xi_n) \leq & 2d_n\mathcal{N}_n \exp\left\{\frac{-n^2M_2^2\xi_n^2}{2nC/h + 2CM_2n\xi_nM_n/(3h)}\right\} \\ \leq & 2d_n\mathcal{N}_n \exp\Big\{ - M_2 \log h^{-1}\Big\} = o(1), \end{aligned} \tag{10.51}$$

where M_2 is chosen such that

$$M_2 > 3C, \quad d_n\mathcal{N}_n \exp\Big\{ - M_2 \log h^{-1}\Big\} = o(1),$$

which are possible as d_n is diverging with certain polynomial rate. Hence we have

$$\Pi_{n3} = O_P(\xi_n). \tag{10.52}$$

In view of (10.45)–(10.48) and (10.52), we have shown (10.44), completing the proof of Lemma 10.1. □

Lemma 10.2. Suppose that Assumptions B1 and B3 in Chapter 10.1 are satisfied. Moreover, suppose that the dimension $d_n \propto \exp\{(nh)^{\tau_2}\}$

with $0 \leq \tau_2 < 1$, $\mathsf{E}\Big\{q_1\big[\sum_{j=1}^{d_n} a_j(U_i)x_{ij}, y_i\big]\big|X_i, U_i\Big\} = 0$ a.s., the moment condition (10.2) holds for all $m \geq 2$, and $h \propto n^{-\delta_1}$ with $0 < \delta_1 < 1$. Then we have, as $n \to \infty$,

$$\max_{1\leq j\leq d_n} \sup_{u\in[0,1]} \Big|\frac{1}{n}\sum_{i=1}^{n} Z_{ij}(u,l)\Big| = o_P\Big(\big(\frac{\log h^{-1}}{nh}\big)^{\tau_3/2}\Big) \tag{10.53}$$

for any $l = 0, 1, 2, \cdots$, $0 < \tau_3 \leq 1 - \tau_2$.

Proof of Lemma 10.2. The proof of (10.53) is similar to the proof of (10.44) in Lemma 10.1. The major difference is the way of dealing with Π_{n1}. Because of the stronger moment condition in (10.2), we may directly use a different exponential inequality and do not need to apply the truncation method. By replacing ξ_n by $\xi_n(\tau_3) \equiv \big(\frac{\log h^{-1}}{nh}\big)^{\tau_3/2}$, we may re-define $r = o\big(\xi_n(\tau_3)h^2\big)$ and thus $\mathcal{N}_n = O\big(r^{-1}\big)$.

Note that there exists a positive constant M_3 such that

$$\mathsf{E}\big[\big|Z_{ij}(u,l)\big|^m\big] \leq \frac{M_3}{2h} m!(h^{-1})^{m-2} \tag{10.54}$$

for all $m \geq 2$, by using the moment condition (10.2). Then, by (10.54) and Lemma 2.2.11 in van der Vaart and Wellner (1996) with $M = h^{-1}$

and $v_i = M_4/h$, we can show that for any $\epsilon > 0$

$$\begin{aligned}
\mathsf{P}\big(\Pi_{n1} > \epsilon\xi_n(\tau_3)\big) \leq & 2d_n\mathcal{N}_n \exp\Big\{\frac{-n^2\epsilon^2\xi_n^2(\tau_3)}{2nM_4/h + 2n\epsilon\xi_n(\tau_3)/h}\Big\} \\
\leq & 2d_n\mathcal{N}_n \exp\Big\{-\frac{\epsilon^2(\log h^{-1})^{\tau_3}}{3M_4}(nh)^{1-\tau_3}\Big\} \\
= & 2\mathcal{N}_n \exp\Big\{(nh)^{\tau_2} - \frac{\epsilon^2\delta_1^{\tau_3}(\log n)^{\tau_3}}{3M_4}(nh)^{1-\tau_3}\Big\} \\
= & o(1) \qquad (10.55)
\end{aligned}$$

as $(1-\tau_3) \geq \tau_2$. The remaining proof is the same as that in the proof of Lemma 10.1. Hence details are omitted here to save space. □

Define

$$\mathcal{M}_{\boldsymbol{\alpha}} = \big(\boldsymbol{\alpha}_j :\ 1 \leq j \leq s_{n2}\big) \ \text{ and } \ \mathcal{M}_{\boldsymbol{\beta}} = \big(\boldsymbol{\beta}_j :\ 1 \leq j \leq s_{n1}\big), \quad (10.56)$$

which correspond the non-zero components in $\mathcal{A}_0$ and $\mathcal{B}_0$, respectively. Let $\dot{\mathcal{L}}_n^{\diamond}(\mathcal{A},\ \mathcal{B} \mid \mathcal{M}_{\boldsymbol{\alpha}})$, $\dot{\mathcal{L}}_n^{\diamond}(\mathcal{A},\ \mathcal{B} \mid \mathcal{M}_{\boldsymbol{\beta}})$, $\dot{\mathcal{L}}_n^{\diamond}(\mathcal{A},\ \mathcal{B} \mid \boldsymbol{\alpha}_j)$ and $\dot{\mathcal{L}}_n^{\diamond}(\mathcal{A},\ \mathcal{B} \mid \boldsymbol{\beta}_j)$ be the gradient vector of $\mathcal{L}_n^{\diamond}(\mathcal{A},\ \mathcal{B})$ with respect to $\mathcal{M}_{\boldsymbol{\alpha}}$, $\mathcal{M}_{\boldsymbol{\beta}}$, $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$, respectively. Define the sub-gradient of the adaptive group LASSO penalty terms in (3.19) as

$$\begin{aligned}
\mathcal{P}^1(\mathcal{M}_{\boldsymbol{\alpha}}) = \Big(&\frac{\alpha_{11}}{\|\widetilde{\boldsymbol{\alpha}}_1\|^{\kappa}\|\boldsymbol{\alpha}_1\|}, \cdots, \frac{\alpha_{s_{n2}1}}{\|\widetilde{\boldsymbol{\alpha}}_{s_{n2}}\|^{\kappa}\|\boldsymbol{\alpha}_{s_{n2}}\|}, \cdots, \\
&\frac{\alpha_{1n}}{\|\widetilde{\boldsymbol{\alpha}}_1\|^{\kappa}\|\boldsymbol{\alpha}_1\|}, \cdots, \frac{\alpha_{s_{n2}n}}{\|\widetilde{\boldsymbol{\alpha}}_{s_{n2}}\|^{\kappa}\|\boldsymbol{\alpha}_{s_{n2}}\|}\Big)^{\mathrm{T}}, \\
\mathcal{P}^1(\mathcal{M}_{\boldsymbol{\beta}}) = \Big(&\frac{\beta_{11}}{\|\widetilde{D}_1\|^{\kappa}\|\boldsymbol{\beta}_1\|}, \cdots, \frac{\beta_{s_{n1}1}}{\|\widetilde{D}_{s_{n1}}\|^{\kappa}\|\boldsymbol{\beta}_{s_{n1}}\|}, \cdots, \\
&\frac{\beta_{1n}}{\|\widetilde{D}_1\|^{\kappa}\|\boldsymbol{\beta}_1\|}, \cdots, \frac{\beta_{s_{n1}n}}{\|\widetilde{D}_{s_{n1}}\|^{\kappa}\|\boldsymbol{\beta}_{s_{n1}}\|}\Big)^{\mathrm{T}}.
\end{aligned}$$

The following lemma is crucial to the proof of Theorem 4.3 (i).

Lemma 10.3. Suppose that the conditions of Theorem 4.3 (i) are satisfied. Then, the objective function $\mathcal{Q}_n^1(\mathcal{A},\ \mathcal{B})$ has a unique maximiser $\big(\widehat{\mathcal{A}}_n^o,\ \widehat{\mathcal{B}}_n^o\big)$ if

$$\dot{\mathcal{L}}_n^\diamond(\mathcal{A},\ \mathcal{B} \mid \mathcal{M}_{\boldsymbol{\alpha}}) - \lambda_3 \mathcal{P}^1(\mathcal{M}_{\boldsymbol{\alpha}}) = \mathbf{0}, \tag{10.57}$$

$$\dot{\mathcal{L}}_n^\diamond(\mathcal{A},\ \mathcal{B} \mid \mathcal{M}_{\boldsymbol{\beta}}) - \lambda_3^* \mathcal{P}^1(\mathcal{M}_{\boldsymbol{\beta}}) = \mathbf{0}, \tag{10.58}$$

$$\max_{s_{n2}+1 \le j \le d_n} \|\dot{\mathcal{L}}_n^\diamond(\mathcal{A},\ \mathcal{B} \mid \boldsymbol{\alpha}_j)\| < \lambda_3 \min_{s_{n2}+1 \le j \le d_n} \|\widetilde{\boldsymbol{\alpha}}_j\|^{-\kappa}, \tag{10.59}$$

$$\max_{s_{n1}+1 \le j \le d_n} \|\dot{\mathcal{L}}_n^\diamond(\mathcal{A},\ \mathcal{B} \mid \boldsymbol{\beta}_j)\| < \lambda_3^* \min_{s_{n1}+1 \le j \le d_n} \|\widetilde{D}_j\|^{-\kappa} \tag{10.60}$$

hold at $\mathcal{A} = \widehat{\mathcal{A}}_n^o$ and $\mathcal{B} = \widehat{\mathcal{B}}_n^o$, where $\mathbf{0}$ is a null vector whose size may change from line to line.

Proof of Lemma 10.3. The proof of this lemma is similar to the proof of Theorem 1 in Fan and Lv (2011). Hence, the details are omitted here to save space. □

Let $\mathcal{P}^2(\mathcal{M}_{\boldsymbol{\alpha}})$ and $\mathcal{P}^2(\mathcal{M}_{\boldsymbol{\beta}})$ be defined as $\mathcal{P}^1(\mathcal{M}_{\boldsymbol{\alpha}})$ and $\mathcal{P}^1(\mathcal{M}_{\boldsymbol{\beta}})$ with $\|\widetilde{\boldsymbol{\alpha}}_j\|^{-\kappa}$ and $\|\widetilde{D}_j\|^{-\kappa}$ being replaced by $\dot{p}_{\lambda_4}(\|\widetilde{\boldsymbol{\alpha}}_j\|)$ and $\dot{p}_{\lambda_4^*}(\|\widetilde{D}_j\|)$, respectively. We next give a lemma for the case of the adaptive SCAD penalty function, which is crucial to the proof of Theorem 4.3 (ii). The proof of Lemma 10.4 below is also similar to the proof of Theorem 1 in Fan and Lv (2011).

Lemma 10.4. Suppose that the conditions of Theorem 4.3 (ii) are satisfied. Then, the objective function $\mathcal{Q}_n^2(\mathcal{A},\ \mathcal{B})$ has a unique maximiser

$\big(\overline{\mathcal{A}}_n^o,\ \overline{\mathcal{B}}_n^o\big)$ if

$$\dot{\mathcal{L}}_n^{\diamond}(\mathcal{A},\ \mathcal{B} \mid \mathcal{M}_{\boldsymbol{\alpha}}) - \mathcal{P}^2(\mathcal{M}_{\boldsymbol{\alpha}}) = \mathbf{0}, \tag{10.61}$$

$$\dot{\mathcal{L}}_n^{\diamond}(\mathcal{A},\ \mathcal{B} \mid \mathcal{M}_{\boldsymbol{\beta}}) - \mathcal{P}^2(\mathcal{M}_{\boldsymbol{\beta}}) = \mathbf{0}, \tag{10.62}$$

$$\max_{s_{n2}+1\le j\le d_n} \|\dot{\mathcal{L}}_n^{\diamond}(\mathcal{A},\ \mathcal{B} \mid \boldsymbol{\alpha}_j)\| < \min_{s_{n2}+1\le j\le d_n} \dot{p}_{\lambda_4}(\|\widetilde{\boldsymbol{\alpha}}_j\|), \tag{10.63}$$

$$\max_{s_{n1}+1\le j\le d_n} \|\dot{\mathcal{L}}_n^{\diamond}(\mathcal{A},\ \mathcal{B} \mid \boldsymbol{\beta}_j)\| < \min_{s_{n1}+1\le j\le d_n} \dot{p}_{\lambda_4^*}(\|\widetilde{D}_j\|) \tag{10.64}$$

hold at $\mathcal{A} = \overline{\mathcal{A}}_n^o$ and $\mathcal{B} = \overline{\mathcal{B}}_n^o$.

Proof of (10.27). Note that $\mathcal{I}_n(1)$ equals to

$$\begin{aligned}
&\mathcal{L}_n^{\diamond}\big(\mathcal{A}_0 + \gamma_n^*\mathcal{U}_1, \mathcal{B}_0 + \gamma_n^*\mathcal{U}_1/h\big) - \mathcal{L}_n^{\diamond}\big(\mathcal{A}_0, \mathcal{B}_0\big)\\
=&\gamma_n^*\mathcal{V}_n^{\mathrm{T}}(\mathcal{U}_1,\ \mathcal{U}_2)\dot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n)\\
&+\frac{1}{2}\Big\{\mathcal{V}_n^{\mathrm{T}}\big(\mathcal{A}_0 - \widetilde{\mathcal{A}}_n + \gamma_n^*\mathcal{U}_1,\ h(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n) + \gamma_n^*\mathcal{U}_2\big)\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n)\\
&\qquad \mathcal{V}_n(\mathcal{A}_0 - \widetilde{\mathcal{A}}_n + \gamma_n^*\mathcal{U}_1,\ h(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n) + \gamma_n^*\mathcal{U}_2)\\
&-\mathcal{V}_n^{\mathrm{T}}(\mathcal{A}_0 - \widetilde{\mathcal{A}}_n,\ h(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n))\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n)\mathcal{V}_n(\mathcal{A}_0 - \widetilde{\mathcal{A}}_n,\ h(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n))\Big\}\\
\equiv&\mathcal{I}_n(1,1) + \mathcal{I}_n(1,2).
\end{aligned}$$

By Taylor's expansion, we have

$$\begin{aligned}
\mathcal{I}_n(1,1) =&\gamma_n^*\mathcal{V}_n^{\mathrm{T}}(\mathcal{U}_1,\ \mathcal{U}_2)\dot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n) \overset{P}{\sim} \gamma_n^*\mathcal{V}_n^{\mathrm{T}}(\mathcal{U}_1,\ \mathcal{U}_2)\dot{\mathcal{L}}_n(\mathcal{A}_0, \mathcal{B}_0)\\
&-\gamma_n^*\mathcal{V}_n^{\mathrm{T}}(\mathcal{U}_1,\ \mathcal{U}_2)\ddot{\mathcal{L}}_n(\mathcal{A}_0, \mathcal{B}_0)\mathcal{V}_n\big(\mathcal{A}_0 - \widetilde{\mathcal{A}}_n,\ h(\mathcal{B}_0 - \widetilde{\mathcal{B}}_n)\big).
\end{aligned}$$

On the other hand, by some elementary calculations, we also have

$$
\begin{aligned}
&\mathcal{I}_n(1,2)\\
=&\frac{1}{2}\Big\{\mathcal{V}_n^{\mathrm{T}}\big(\mathcal{A}_0-\widetilde{\mathcal{A}}_n+\gamma_n^*\mathcal{U}_1,\ h(\mathcal{B}_0-\widetilde{\mathcal{B}}_n)+\gamma_n^*\mathcal{U}_2\big)\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n)\mathcal{V}_n\big(\mathcal{A}_0-\widetilde{\mathcal{A}}_n+\gamma_n^*\mathcal{U}_1,\ h(\mathcal{B}_0-\widetilde{\mathcal{B}}_n)+\gamma_n^*\mathcal{U}_2\big)\\
&-\mathcal{V}_n^{\mathrm{T}}\big(\mathcal{A}_0-\widetilde{\mathcal{A}}_n,\ h(\mathcal{B}_0-\widetilde{\mathcal{B}}_n)\big)\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n)\mathcal{V}_n\big(\mathcal{A}_0-\widetilde{\mathcal{A}}_n+\gamma_n^*\mathcal{U}_1,\ h(\mathcal{B}_0-\widetilde{\mathcal{B}}_n)+\gamma_n^*\mathcal{U}_2\big)\\
&+\mathcal{V}_n^{\mathrm{T}}\big(\mathcal{A}_0-\widetilde{\mathcal{A}}_n,\ h(\mathcal{B}_0-\widetilde{\mathcal{B}}_n)\big)\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n)\mathcal{V}_n\big(\mathcal{A}_0-\widetilde{\mathcal{A}}_n+\gamma_n^*\mathcal{U}_1,\ h(\mathcal{B}_0-\widetilde{\mathcal{B}}_n)+\gamma_n^*\mathcal{U}_2\big)\\
&-\mathcal{V}_n^{\mathrm{T}}\big(\mathcal{A}_0-\widetilde{\mathcal{A}}_n,\ h(\mathcal{B}_0-\widetilde{\mathcal{B}}_n)\big)\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\ \widetilde{\mathcal{B}}_n)\mathcal{V}_n\big(\mathcal{A}_0-\widetilde{\mathcal{A}}_n,\ h(\mathcal{B}_0-\widetilde{\mathcal{B}}_n)\big)\Big\}\\
=&\frac{\gamma_n^*}{2}\mathcal{V}_n^{\mathrm{T}}(\mathcal{U}_1,\ \mathcal{U}_2)\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\widetilde{\mathcal{B}}_n)\mathcal{V}_n\big(\mathcal{A}_0-\widetilde{\mathcal{A}}_n+\gamma_n^*\mathcal{U}_1,\ h(\mathcal{B}_0-\widetilde{\mathcal{B}}_n)+\gamma_n^*\mathcal{U}_2\big)\\
&+\frac{\gamma_n^*}{2}\mathcal{V}_n\big(\mathcal{A}_0-\widetilde{\mathcal{A}}_n,\ h(\mathcal{B}_0-\widetilde{\mathcal{B}}_n)\big)\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\widetilde{\mathcal{B}}_n)\mathcal{V}_n(\mathcal{U}_1,\ \mathcal{U}_2)\\
=&\frac{1}{2}(\gamma_n^*)^2\mathcal{V}_n^{\mathrm{T}}(\mathcal{U}_1,\ \mathcal{U}_2)\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\widetilde{\mathcal{B}}_n)\mathcal{V}_n(\mathcal{U}_1,\ \mathcal{U}_2)+\gamma_n^*\mathcal{V}_n\big(\mathcal{A}_0-\widetilde{\mathcal{A}}_n,\ h(\mathcal{B}_0-\widetilde{\mathcal{B}}_n)\big)\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\widetilde{\mathcal{B}}_n)\mathcal{V}_n(\mathcal{U}_1,\ \mathcal{U}_2)\\
\overset{P}{\sim}&\frac{1}{2}(\gamma_n^*)^2\mathcal{V}_n^{\mathrm{T}}(\mathcal{U}_1,\ \mathcal{U}_2)\ddot{\mathcal{L}}_n(\widetilde{\mathcal{A}}_n,\widetilde{\mathcal{B}}_n)\mathcal{V}_n(\mathcal{U}_1,\ \mathcal{U}_2)+\gamma_n^*\mathcal{V}_n^{\mathrm{T}}(\mathcal{U}_1,\ \mathcal{U}_2)\ddot{\mathcal{L}}_n(\mathcal{A}_0,\mathcal{B}_0)\mathcal{V}_n\big(\mathcal{A}_0-\widetilde{\mathcal{A}}_n,\ h(\mathcal{B}_0-\widetilde{\mathcal{B}}_n)\big).
\end{aligned}
$$

We can easily prove (10.27) by using the above two results on asymptotic expansion for $\mathcal{I}_n(1,1)$ and $\mathcal{I}_n(1,2)$. □

Proof (10.28). Recall that

$$\mathcal{I}_n(4)=\gamma_n^*\mathcal{V}_n^{\mathrm{T}}(\mathcal{U}_1,\mathcal{U}_2)\dot{\mathcal{L}}_n(\mathcal{A}_0,\mathcal{B}_0). \tag{10.65}$$

By Taylor's expansion for $q_1(r,y)$ and Assumption B4, we have

$$
\begin{aligned}
&q_1\Big\{\sum_{j=1}^{d_n}\big[a_j(U_k)+\dot{a}_j(U_k)(U_i-U_k)\big]x_{ij},\ y_i\Big\}\\
=&q_1\Big\{\sum_{j=1}^{s_{n2}}\big[a_j(U_k)+\dot{a}_j(U_k)(U_i-U_k)\big]x_{ij},\ y_i\Big\}\\
=&q_1\Big[\sum_{j=1}^{s_{n2}}a_j(U_i)x_{ij},\ y_i\Big]+O_P(s_{n2}h^2),
\end{aligned}
\tag{10.66}
$$

which implies that

$$\begin{aligned}\mathcal{I}_n(4) =& \frac{\gamma_n^*}{n}\sum_{k=1}^{n}\sum_{i=1}^{n} q_1\big[\sum_{j=1}^{s_{n2}} a_j(U_i)x_{ij},\ y_i\big]X_i^{\mathrm{T}}\mathbf{u}_1(k)K_h(U_i-U_k)\\ &+\frac{\gamma_n^*}{n}\sum_{k=1}^{n}\sum_{i=1}^{n} q_1\Big[\sum_{j=1}^{s_{n2}} a_j(U_i)x_{ij},\ y_i\Big]\Big(\frac{U_i-U_k}{h}\Big)X_i^{\mathrm{T}}\mathbf{u}_2(k)K_h(U_i-U_k)\\ &+O_P(\gamma_n^* s_{n2}^{3/2}n^{1/2}h^2)\cdot\big(\|\mathcal{U}_1\|+\|\mathcal{U}_2\|\big). \qquad (10.67)\end{aligned}$$

Note that (U_i, X_i, y_i), $i = 1, \cdots, n$, are independent and identically distributed. By Assumptions B1, B2(i) and B3 in Chapter 10.1, and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\mathsf{E}\Big[\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n} q_1\big[\sum_{j=1}^{s_{n2}} a_j(U_i)x_{ij},\ y_i\big]X_i^{\mathrm{T}}\mathbf{u}_1(k)K_h(U_i-U_k)\Big]^2\\ \le&\frac{1}{n}\sum_{k=1}^{n}\mathsf{E}\Big\{\sum_{i=1}^{n} q_1\big[\sum_{j=1}^{s_{n2}} a_j(U_i)x_{ij},\ y_i\big]X_i^{\mathrm{T}}\mathbf{u}_1(k)K_h(U_i-U_k)\Big\}^2\\ =&\frac{1}{n}\sum_{k=1}^{n}\mathsf{E}\Big[\mathsf{E}\Big(\Big\{\sum_{i=1}^{n} q_1\big[\sum_{j=1}^{s_{n2}} a_j(U_i)x_{ij},\ y_i\big]X_i^{\mathrm{T}}\mathbf{u}_1(k)K_h(U_i-U_k)\Big\}^2\Big|U_k\Big)\Big]\\ =&\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n}\mathsf{E}\Big[\mathsf{E}\Big(\Big\{q_1^2\big[\sum_{j=1}^{s_{n2}} a_j(U_i)x_{ij},\ y_i\big]\Big\}\mathbf{u}_1^{\mathrm{T}}(k)X_iX_i^{\mathrm{T}}\mathbf{u}_1(k)K_h^2(U_i-U_k)\Big|U_k\Big)\Big]\\ =&O\Big(h^{-1}\sum_{k=1}^{n}\mathbf{u}_1^{\mathrm{T}}(k)\mathbf{u}_1(k)\Big)=O(h^{-1})\cdot\|\mathcal{U}_1\|^2.\end{aligned}$$

Similarly, we can also show that

$$\begin{aligned}&\mathsf{E}\big[\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n} q_1\Big[\sum_{j=1}^{s_{n2}} a_j(U_i)x_{ij},\ y_i\Big]\Big(\frac{U_i-U_k}{h}\Big)X_i^{\mathrm{T}}\mathbf{u}_2(k)K_h(U_i-U_k)\Big]^2\\ &=O(h^{-1})\cdot\|\mathcal{U}_2\|^2.\end{aligned}$$

Noting that $s_{n2}h^2 \propto (nh)^{-1/2}$, we have

$$\mathcal{I}_n(4) = O_P\big((\gamma_n^*)^2 n^{1/2}\big) \cdot \big(\|\mathcal{U}_1\| + \|\mathcal{U}_2\|\big), \tag{10.68}$$

which completes the proof of (10.28). □

Keywords and phrases: GSVCM, LASSO, SCAD, local maximum likelihood, penalise likelihood, model selection, oracle estimation, sparsity, ultra-high dimension, prognostic classification

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood pronciple. *2nd Internatinal Symposium on Information Theory*, 267-281, Budapest.

Antoniadis, A. and Fan, J. (2001). Regularized wavelet approximations (with discussion). *Journal of American Statistical Association*, **96**, 939-967.

Bickel, P. J. (1975). One-step Huber estimates in linear models. *Journal of the American Statistical Association* **70**, 428–433.

Bickel, P. J., 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.

Breiman, L. (1996). Heuristics of Instability and Stabilization in Model Selection. *The Annals of Statistics*, **24**, 2350-2383.

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

Cleveland, W. S., Grosse, E. and Shyu, W. M. (1991). Local regression models. In *Statistical Models in S* (eds J. M. Chambersand T. Hastie), pp. 309-376. Pacific Grove: Wadsworthand Brooks/Cole

Combe B., Cantagrel A., Goupille P., et al. (2003). Predictive factors of 5-year health assessment questionnaire disability in early rheumatoid arthritis. *The Journal of Rheumatology* , **30**, 2344-9.

Combe, B., Landewe, R., Lukas, C., et al. (2007). EULAR evidence recommendations for the management of early arthritis. Report of a task force of the European Standing Committee for International Clinical Studies Including Therapeutics. *Annals of the Rheumatic Diseases*, **66**, 34-45.

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.

Dixon, N. and Symmons, D. (2005). Does early rheumatoid arthritis exist? *Best Practice & Research Clinical Rheumatology*, **19**, 37-53.

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. (1997). "Comments on 'Wavelets in Statistics: A Review' by A. Antoniadis." *Journal of the Italian Statistical Association*, **6**, 131-138.

Fan, J. and Chen, J. (1999). One-step local quasi-likelihood estimation. *Journal of the Royal Statistical Society Series B*, **61**, 927–943.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society Series B*, **57**, 371 – 394.

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 Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, **30**, 74–99.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, **99**, 710–723.

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. (2010). A selection overview of variable selection in high dimensional feature space. *Statistica Sinica*, **20**, 101–148.

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. Forthcoming in *Journal of the American Statistical Association*.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with diverging number of parameters. *The Annals of Statistics*, **32**, 928–961.

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. 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, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and its Interface*, **1**(1), 179-195.

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.

Fang, X., Li, J., Wong, W. K., and Fu, B. (2014). Detecting the violation of variance homogeneity in mixed models. *Statistical methods in medical research*, 0962280214526194.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**(2), 109-135.

Farragher, T., Lunt, M., Fu, B., Bunn, D. and Symmons, D. (2010). Early treatment with, and time receiving, first disease-modifying antirheumatic drug predicts long-term function in patients with

inflammatory polyarthritis. *Annals of Rheumatic Diseases*, **69**(4), 689-695.

Fu, B., Lunt, M., Galloway, J., Dixon, W., Hyrich, K., and Symmons, D. (2013). A Threshold Hazard Model for Estimating Serious Infection Risk Following Anti-Tumor Necrosis Factor Therapy in Rheumatoid Arthritis Patients. *Journal of Biopharmaceutical Statistics*, **23**(2), 461-476.

Greenland, S. (2007). Bayesian perspectives for epidemiological research. II. Regression analysis. *International Journal of Epidemiology*, **36**(1), 195-202.

Hastie, T. J. and Tibshirani, R. J. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society Series B*, **55**, 757–796.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**(1), 55–67.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley & Sons.

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.

Jiang, J. (2014). Multivariate Functional-coefficient Regression Models for Multivariate Nonlinear Times Series. *Biometrika*, doi: 10.1093/biomet/asu011.

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.

Ke, Y., Fu, B. and Zhang, W. (2015). A semi-varying coefficient multinomial logistic regression for prognostic classification with application to stratified medicine. Working paper, Department of Mathematics, University of York.

Kong, E. and Xia, Y. (2006) Variable selection for the single-index model. *Biometrika*, **94**, 217-229.

Kuk, A. Y. C., Li, J. and Rush, J. A. (2014). Variable and threshold selection to control predictive accuracy in logistic regression. *Applied Statistics*, DOI: 10.1111/rssc.12058.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, **22**, 4555

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). Model selection and structure specification in ultra-high dimensional generalised semi-varying coefficient models. Working paper, Department of Mathematics, University of York.

Li, J., Jiang, B. and Fine, J.(2013). Multicategory reclassification statistics for assessing improvements in diagnostic accuracy. *Biostatistics*. 14(2): 382-394.

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.

Mack, Y. P. and Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift fur Wahrscheinlichskeittheorie und verwandte Gebiete*, **61**, 405–415.

Mosteller, F. and Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *Journal of the American Statistical Association*, **58**(302), 275-309.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, **7**, 186–199.

Ruppert, D. and Wand, M.P. (1994). Multivariate weighted least squares regression. *The Annals of Statistics*, **22**, 1346-1370.

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461-464.

Solari, A., Saskia C. and Jelle J. G. (2012). Testing goodness of fit in regression: a general approach for specified alternatives. *Statistics in medicine*, **31**(28), 3656-3666.

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

Stefanski, L., Wu, Y., and White, K. (2014). Variable selection in nonparametric classification via measurement error model selection likelihoods. *Journal of the American Statistical Association*, DOI:10.1080/01621459.2013.858630.

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.

Sun, Y., Zhang, W. and Tong, H. (2007). Estimation of the covariance matrix of random effects in longitudinal studies. *The Annals of Statistics*, 2795-2814.

Tao, H. and Xia, Y. (2011) Adaptive Semi-varying Coefficient Model Selection, *Statistica Sinica*, **22**, 575-599.

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

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the Lasso. *The Annals of Statistics*, **36** , 614-645.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes with applications to statistics*. Springer Series in Statistics, Springer.

Venkateshan, S., Sidhu, S., Malhotra, S., et al. (2009). Efficacy of biologicals in the treatment of rheumatoid arthritis: a meta-analysis. *Pharmacology*, **83**, 1-9.

Visser, H. (2005). Early diagnosis of rheumatoid arthritis. *Best Practice & Research Clinical Rheumatology*, **19**, 55-72.

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.

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

Wu, Y., Fan, J., and Muller, H. (2010). Varying-coefficient functional linear regression. *Bernoulli*, **16**, 730-758.

Xia, Y., Zhang, W. and Tong, H. (2004). Efficient estimation for semivarying coefficient models. *Biometrika*, **91**, 661-681.

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. (2011). Identification of the constant components in generalised semivarying coefficient models by cross-validation. *Statistica Sinica*, **21**, 1913–1929.

Zhang, W., Lee, S. Y. and Song, X. (2002). Local polynomial fitting in semivarying coefficient models. *Journal of Multivariate Analysis*, **82**, 166-188.

Zhang, W., Li, D. and Xia, Y. (2012). An iterative estimation procedure for generalised varying-coefficient models with unspecified link functions. Manuscript.

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 Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, **67**(2), 301–320.

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.