International Statistical Review

International Statistical Review (2015), 83, 1, 36–64 doi:10.1111/insr.12029

Varying Coefficient Regression Models: A Review and New Developments[1]

Byeong U. Park[1], Enno Mammen[2], Young K. Lee[3] and Eun Ryung Lee[2]

[1]*Seoul National University, Seoul, Korea*
E-mail: bupark@stats.snu.ac.kr
[2]*Universität Mannheim, Mannheim, Germany*
E-mail: emammen@rumms.uni-mannheim.de, E-mail: silverryuee@gmail.com
[3]*Kangwon National University, Chuncheon, Korea*
E-mail: youngklee@kangwon.ac.kr

Summary

Varying coefficient regression models are known to be very useful tools for analysing the relation between a response and a group of covariates. Their structure and interpretability are similar to those for the traditional linear regression model, but they are more flexible because of the infinite dimensionality of the corresponding parameter spaces. The aims of this paper are to give an overview on the existing methodological and theoretical developments for varying coefficient models and to discuss their extensions with some new developments. The new developments enable us to use different amount of smoothing for estimating different component functions in the models. They are for a flexible form of varying coefficient models that requires smoothing across different covariates' spaces and are based on the smooth backfitting technique that is admitted as a powerful technique for fitting structural regression models and is also known to free us from the curse of dimensionality.

Key words: Varying coefficient models; kernel smoothing; sieve estimation; penalised likelihood methods; partially linear models; longitudinal data; projection; quasi-likelihood; integral equation; shrinkage estimation; robust estimation.

1 Introduction

It is widely known that nonparametric methods fail when they are applied to high-dimensional spaces. Structural nonparametric regression is one way of avoiding the curse of dimensionality. Two useful examples of structural regression are additive models introduced by Breiman & Friedman (1985) and varying coefficient models proposed by Hastie & Tibshirani (1993). In additive models, the regression function is expressed as a sum of univariate functions of covariates. In varying coefficient models, unlike the classical linear regression models, the regression coefficients are not set to be constants but are allowed to depend on

[1] This paper is followed by discussions and a rejoinder.

some other covariate(s). As we will see in the following text, a general form of varying coefficient models includes additive models as special cases. Thus, we focus on varying coefficient models in this paper.

Varying coefficient models inherit simplicity and easy interpretation of the traditional linear models, yet are intrinsically nonparametric. They arise in many real applications, see Hastie & Tibshirani (1993) and Fan & Zhang (2008) for various applications of the models. In this paper, we present an overview on earlier methodological and theoretical developments for varying coefficient models and also discuss their extensions with some new developments. Our main focus is on the kernel smoothing technique, although we also discuss sieve and penalised likelihood methods. The former is methodologically more challenging for varying coefficient models.

There are two possibilities in building a varying coefficient model. One is to let all regression coefficients depend on a single covariate. With this option, the mean regression function takes the form

$$E(Y|\mathbf{X} = \mathbf{x}, Z = z) = x_1 f_1(z) + \cdots + x_d f_d(z), \tag{1.1}$$

where Y is a response variable, $\mathbf{X} = (X_1, \ldots, X_d)^\top$ and Z are covariates and f_j are the unknown coefficient functions. The other is to let f_j for different j be functions of different covariates, which leads to the model

$$E(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = x_1 f_1(z_1) + \cdots + x_d f_d(z_d), \tag{1.2}$$

where $\mathbf{Z} = (Z_1, \ldots, Z_d)^\top$ is the vector of the covariates that modify the effects of the covariates X_j in nonparametric ways. Throughout this paper, we call Z_j 'smoothing variables'. There is a big difference between fitting the two models (1.1) and (1.2) by the kernel smoothing technique. For (1.1), a pointwise kernel-weighted least squares fit gives directly estimators of the coefficient functions f_j, whereas for (1.2), it does not give proper estimators. For a more detailed account of this issue, see Section 3.1.

There are various extensions of models (1.1) and (1.2), including those that accommodate discrete responses and those where a covariate may take both the roles of linear effect variables (X_j) and of smoothing variables (Z_j). These extended models give more flexibility in applying varying coefficient models. One may also think of variants of models (1.1) and (1.2) for longitudinal data. Partially linear models where f_j are constants for some j are other options. We will discuss the methodology and theory for fitting these models. We will also extend the discussion to other related problems such as construction of confidence intervals, hypothesis testing, quantile estimation, bandwidth selection and variable selection for sparse models.

Suppose we fit model (1.1) using the kernel smoothing technique. Because this model has a common smoothing variable, Z, and kernel smoothing acts on the space of the smoothing variable, we would take the local least squares fit that uses a single bandwidth. But this may not give sufficiently good estimators because the functions f_j may have different shapes and thus need different amount of smoothing on the space of the smoothing variable Z. In fact, Fan & Zhang (1999) suggested an idea of two-step estimation. We elaborate this idea further with a slight modification in Section 2.3. Furthermore, we extend the idea to model (1.2) and its generalisations in Section 3.3. In the latter, we use the 'smooth backfitting' technique (Mammen *et al.*, 1999; Lee *et al.*, 2012a) in the first-step estimation. We provide some numerical evidences of the two-step procedures as well as their theoretical properties. These are new in this paper.

On the following, Sections 2–4 are focused on the kernel smoothing technique, Section 5 is devoted to the discussion of sieve and penalised likelihood estimation and in Section 6, a real dataset is analysed through generalised versions of model (1.2).

2 Kernel Estimation: Single Smoothing Variable

There exists a large body of literature working on kernel methods for fitting model (1.1), where all coefficient functions are defined on the space of a single 'smoothing' (univariate or multivariate) variable Z. In this model, each X_j may be supported on an interval or on a finite set. Fitting model (1.1) is simple. A standard kernel smoothing across the single variable Z gives directly proper estimators of f_j. In Section 2.1, we discuss kernel estimation of model (1.1). In Section 2.2, we introduce some existing results and possible extensions for (1.1) and related models. In Section 2.3, we get into details about a two-step approach to the estimation of model (1.1) and provide some new results.

2.1 Basic Methods and Theory

The Nadaraya–Watson estimators of $f_j(z)$ for $1 \le j \le d$ are obtained by minimising

$$\sum_{i=1}^{n} \left(Y^i - \sum_{j=1}^{d} a_j X_j^i \right)^2 K_h\left(z, Z^i\right) \tag{2.1}$$

with respect to a_j, where $K_h(z,u) = K((z-u)/h)/h$, h is the bandwidth, and K is the nonnegative kernel function with $\int K = 1$. We denote the estimators by $\hat{f}_j$. Let $\mathbb{X}$ denote the $(n \times d)$ matrix such that X_j^i is its (i,j)-th entry, $\mathbb{W}(z)$ denote the $(n \times n)$ diagonal matrix with $K_h(z, Z^i)$ being its diagonal entries and $\mathbb{Z} = (Z^1, \ldots, Z^n)^\top$. If we write $\mathbb{Y} = (Y^1, \ldots, Y^n)^\top$ and $\hat{\mathbf{f}}(z) = (\hat{f}_1(z), \ldots, \hat{f}_d(z))^\top$, then $\hat{\mathbf{f}}(z) = [\mathbb{X}^\top \mathbb{W}(z)\mathbb{X}]^{-1}\mathbb{X}^\top \mathbb{W}(z)\mathbb{Y}$.

The local linear smoothing technique minimises

$$\sum_{i=1}^{n} \left[Y^i - \sum_{j=1}^{d} \left(a_{0j} + a_{1j}\left(Z^i - z\right)\right) X_j^i \right]^2 K_h\left(z, Z^i\right)$$

with respect to (a_{0j}, a_{1j}). With a slight abuse of notation, let $\hat{f}_j$ denote the local linear estimator of f_j. Let $\mathbb{X}(z)$ be the $(n \times 2d)$ matrix whose (i,j)-th and $(i, d+j)$-th entries for $1 \le j \le d$ are X_j^i and $(Z^i - z)X_j^i$, respectively. Then, $\hat{\mathbf{f}}(z) = (\hat{f}_1(z), \ldots, \hat{f}_d(z))^\top$ is given by

$$\hat{\mathbf{f}}(z) = (\mathbf{I}_d, \mathbf{O}_d)\left[\mathbb{X}(z)^\top \mathbb{W}(z)\mathbb{X}(z)\right]^{-1} \mathbb{X}(z)^\top \mathbb{W}(z)\mathbb{Y},$$

where $\mathbf{I}_d$ and $\mathbf{O}_d$, respectively, denote the identity and zero matrices of dimension d.

The asymptotic distributions of $\hat{f}_j$ can be derived by the standard kernel smoothing theory. Both the Nadaraya–Watson and the local linear estimators have asymptotically normal distributions. Their asymptotic variances are identical. Let $\mathbf{N}(z) = (N_{jk}(z))_{d \times d}$, where $N_{jk}(z) = E(X_j X_k | Z = z)$ and assume for simplicity $\mathrm{Var}(Y|\mathbf{X}, Z) = \sigma^2(Z)$. Let p_0 and p denote the density functions of Z and $(\mathbf{X}, Z)$, respectively. Then, the common asymptotic variance is equal to $n^{-1}h^{-1}\mathbf{N}(z)^{-1}\sigma^2(z)p_0(z)^{-1}\int K^2$. Their asymptotic biases take the form $h^2\mathbf{b}(z)\int u^2 K/2$ with appropriate definitions of $\mathbf{b}(z)$. For the Nadaraya–Watson, $\mathbf{b}(z) = \mathbf{f}''(z) + 2\,\mathbf{N}(z)^{-1}E[\mathbf{X}\mathbf{X}^\top p^{(1)}(\mathbf{X}, Z)/p(\mathbf{X}, Z)]\mathbf{f}'(z)$, where $\mathbf{f}'(z) = (f_1'(z), \ldots, f_d'(z))^\top$, $\mathbf{f}''(z) = (f_1''(z), \ldots, f_d''(z))^\top$ and $p^{(1)}(\mathbf{x}, z) = \partial p(\mathbf{x}, z)/\partial z$. For the local linear, $\mathbf{b}(z)$ is simply equal to $\mathbf{f}''(z)$.

We remark that the covariances between $\hat{f}_j$ and $\hat{f}_k$ for different $j \neq k$ are asymptotically not negligible because $\mathbf{N}(z)$ is not a diagonal matrix. This is not the case with the kernel estimators for the varying coefficient model (1.2), where the coefficients f_j are functions of different smoothing variables. In the latter case, the asymptotic covariances between the kernel estimators of different coefficient functions are negligible, see our discussion in Section 3. We also note that the aforementioned results can be easily extended to the case of a single multivariate smoothing variable $\mathbf{Z}$, where the mean function $m(\mathbf{x}, \mathbf{z}) = E(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$ is expressed as

$$m(\mathbf{x}, \mathbf{z}) = x_1 f_1(\mathbf{z}) + \cdots + x_d f_d(\mathbf{z}). \tag{2.2}$$

Estimation of the latter model (2.2), however, suffers from the curse of dimensionality as the dimension of $\mathbf{Z}$ becomes high.

2.2 Related Problems

A closer look at the estimation procedures in Section 2.1 reveals that one uses the same bandwidth h and kernel K for all coefficient functions f_j. If one wants to use different bandwidths for different functions, one may implement the idea of two-step estimation studied by Fan & Zhang (1999) and Cai (2002). In the next separate section, we will review this issue and develop some new results. In this section, we review some other earlier developments and discuss their possible extensions.

The estimation method for model (1.1) may be extended to the generalised varying coefficient model

$$g(m(\mathbf{x}, z)) = x_1 f_1(z) + \cdots + x_d f_d(z), \tag{2.3}$$

where g is a link function that enables one to apply the model to discrete responses. This was performed by Kauermann & Tutz (1999), Kauermann & Tutz (2000) and Cai (2000). In particular, Cai *et al.* (2000) discussed a goodness-of-fit test to detect whether certain coefficient functions f_j in (2.3) are constant. Fan & Zhang (2000) considered the construction of simultaneous confidence bands for f_j in model (1.1) as well as testing problems whether some coefficient functions f_j belong to a parametric family. Recently, Zhang & Peng (2010) extended the work of Fan & Zhang (2000) to the generalised model (2.3). Galindo *et al.* (2001) also worked on the generalised model (2.3) in a framework of estimating equation and proposed a bootstrap method to construct pointwise confidence intervals for the true coefficient functions.

Selection of smoothing parameters is an important problem in nonparametric function estimation. Two general approaches are cross-validation and plug-in. Cross-validation is computationally expensive, whereas plug-in relies too much on asymptotic analysis. For the particular problem of estimating model (1.1), one may select h that minimises a suitable estimator of the conditional mean squared error of the estimated model,

$$\begin{aligned}\mathrm{MSE}(h) &= \int E\left[\left(\mathbf{x}^\top\left(\hat{\mathbf{f}}(z) - \mathbf{f}(z)\right)\right)^2 \middle| \mathbb{X}, \mathbb{Z}\right] P(d\mathbf{x}, dz) \\ &= \int \left[tr\left(\mathbf{N}(z) \mathrm{V}ar(\hat{\mathbf{f}}(z)|\mathbb{X}, \mathbb{Z})\right) + \mathrm{bias}\left(\hat{\mathbf{f}}(z)|\mathbb{X}, \mathbb{Z}\right)^\top \mathbf{N}(z)\, \mathrm{bias}\left(\hat{\mathbf{f}}(z)|\mathbb{X}, \mathbb{Z}\right)\right] p_0(z)\, dz,\end{aligned}$$

where P denotes the joint distribution of $(\mathbf{X}, Z)$. Zhang & Lee (2000) and Fan & Zhang (2008) took this approach. In particular, Fan & Zhang (2008) proposed a bandwidth selector that minimises the estimated criterion

$$n^{-1} \sum_{i=1}^{n} \left[\mathrm{tr}\left(\mathbf{N}\left(Z^i\right) \hat{\mathbf{V}}_{-i}\left(Z^i\right)\right) + \hat{\mathbf{B}}_{-i}\left(Z^i\right)^\top \mathbf{N}\left(Z^i\right) \hat{\mathbf{B}}_{-i}\left(Z^i\right)\right],$$

where $\hat{\mathbf{V}}_{-i}(z)$ and $\mathbf{B}_{-i}(z)$, respectively, are leave-one-out estimates of $\mathrm{Var}(\hat{\mathbf{f}}(z)|\mathbb{X},\mathbb{Z})$ and $\mathrm{bias}(\hat{\mathbf{f}}(z)|\mathbb{X},\mathbb{Z})$ that are constructed on the basis of $\{(\mathbf{X}^l, Z^l, Y^l) : l \neq i\}$. An alternative approach is based on a penalised sum of squared residuals. In classical nonparametric regression, Härdle *et al.* (1998) showed that a penalised sum of squared residuals is asymptotically equivalent to cross-validation. Penalised least squares bandwidth selection is computationally more feasible than cross-validation. The technique was elaborated recently by Mammen & Park (2005) for the smooth backfitting estimators of the additive regression model. It can be also adapted to the estimation of the varying coefficient models (1.1) and (2.3).

Another interesting problem is to estimate model (1.1) when the covariates X_j have measurement errors. You *et al.* (2006) studied this problem. Suppose that we observe $\tilde{\mathbf{X}}^i = \mathbf{X}^i + \boldsymbol{\zeta}^i$, instead of $\mathbf{X}^i$, where $\boldsymbol{\zeta}^i$ are i.i.d. with mean zero, independent of $(\mathbf{X}^i, Z^i)$ and have a known variance $\boldsymbol{\Sigma}$. In this case, the kernel estimators $\hat{f}_j$ with the contaminated covariates $\tilde{X}^i_j$ leads to inconsistent estimators. To see this, in the case of Nadaraya–Watson smoothing, note that

$$E\left(\hat{\mathbf{f}}(z)|\mathbb{X},\mathbb{Z}\right) = \left[\mathbb{X}^\top \mathbb{W}(z)\mathbb{X} + n\boldsymbol{\Sigma}\right]^{-1} \mathbb{X}^\top \mathbb{W}(z)\mathbb{X}\,\mathbf{f}(z) + o_p(1).$$

To remedy the inconsistency, You *et al.* (2006) took $\left[\tilde{\mathbb{X}}^\top \mathbb{W}(z)\tilde{\mathbb{X}} - n\boldsymbol{\Sigma}\right]^{-1} \tilde{\mathbb{X}}^\top \mathbb{W}(z)\mathbb{Y}$ as an estimator of $\mathbf{f}(z)$ and derived its asymptotic distribution. In case $\boldsymbol{\Sigma}$ is unknown, one may estimate $\boldsymbol{\Sigma}$ based on repeated measurements of $\tilde{\mathbf{X}}^i$.

Robustness has been a favourite subject in most areas of statistical inference. The varying coefficient model is no exception. Indeed, Tang & Wang (2005) considered the varying coefficient model for median regression,

$$\mathrm{median}(Y|\mathbf{X} = \mathbf{x}, Z = z) = x_1 f_1(z) + \cdots + x_d f_d(z).$$

They studied the local linear absolute deviation method, which minimises

$$\sum_{i=1}^{n} \left| Y^i - \sum_{j=1}^{d} \left(a_{0j} + a_{1j}\left(Z^i - z\right)\right) X^i_j \right| K_h\left(z, Z^i\right).$$

Honda (2004) also studied the estimation of the conditional median and discussed briefly its extension to conditional α-quantiles for $\alpha \neq 1/2$. More recently, Wang *et al.* (2009) worked on the local linear Wilcoxon rank estimators, which minimise the local linear rank objective function defined by

$$n^{-1}(n-1)^{-1} \sum_{1 \le i,j \le n} \left|e^i - e^j\right| K_h\left(z, Z^i\right) K_h\left(z, Z^j\right),$$

where $e^i = Y^i - \sum_{j=1}^{d}\left(a_{0j} + a_{1j}(Z^i - z)\right) X^i_j$. In the absence of the kernel weights, minimising the objective function leads to the classical Wilcoxon rank estimators in linear models.

One practically useful variant of model (1.1) is the partially linear varying coefficient model,

$$m(\mathbf{x}, \mathbf{u}, z) = \mathbf{f}(z)^\top \mathbf{x} + \boldsymbol{\beta}^\top \mathbf{u}, \tag{2.4}$$

where $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_q\right)^\top$ is a q-dimensional vector of unknown parameters. One may apply the profiling technique to this model to estimate the unknown parameters β_j and coefficient functions f_j. The resulting estimators of β_j and f_j have an explicit form, and their theoretical properties can be derived. Indeed, Fan & Huang (2005) obtained the asymptotic distributions of the estimators and also considered testing problems for $\boldsymbol{\beta}$ and $\mathbf{f}$. You & Chen (2006) and Zhou & Liang (2009) extended the work of Fan & Huang (2005) to the case where all or some of the linear covariates U_j^i are subject to error. Kai *et al.* (2011) took a similar approach, but they worked on quantile regression models. They also proposed an approach to the estimation of β_j and f_j in the mean regression model (2.4) via quantile regression.

Recent years have seen many proposals for shrinkage or regularised estimation, especially in the setting of the classical linear regression models. Various shrinkage methods have been developed. Two most popular are the LASSO (Tibshirani, 1996; Zou, 2006) and the SCAD (Fan & Li, 2001). Most of the works have been for parametric models or for parametric parts in semiparametric models. There have been a few attempts to extend the idea of shrinkage estimation to kernel smoothing. For the varying coefficient model (1.1), in particular, we are only aware of the works by Wang & Xia (2009) and Hu & Xia (2012). Suppose that f_j is identically zero for all j in some $\mathcal{I}_0 \subset \{1, 2, \ldots, d\}$. Combining the group LASSO idea of Yuan & Lin (2006) and the Nadaraya–Watson smoothing, one may minimise

$$\sum_{i=1}^{n} \sum_{i'=1}^{n} \left(Y^i - \sum_{j=1}^{d} \beta_{j(i')} X_j^i \right)^2 K_h\left(Z^{i'}, Z^i\right) + \sum_{j=1}^{d} \lambda_j \|\boldsymbol{\beta}_j\|, \tag{2.5}$$

where $\boldsymbol{\beta}_j = \left(\beta_{j(1)}, \ldots, \beta_{j(n)}\right)^\top$ and $\|\cdot\|$ denote the Euclidean norm. The minimiser $\hat{\boldsymbol{\beta}}_j$ estimates $\left(f_j(Z^1), \ldots, f_j(Z^n)\right)^\top$. Wang & Xia (2009) claimed that the method correctly identifies $\mathcal{I}_0$ with probability tending to one, and that the resulting estimators are as efficient as the oracle estimators, which uses the knowledge of $\mathcal{I}_0$. Hu & Xia (2012) treated the case where some of f_j are constant.

The variable selection idea of Wang & Xia (2009) may be applied to a general penalty scheme. Instead of (2.5), one may minimise

$$\sum_{i=1}^{n} \sum_{i'=1}^{n} \left(Y^i - \sum_{j=1}^{d} \beta_{j(i')} X_j^i \right)^2 K_h\left(Z^{i'}, Z^i\right) + \sum_{j=1}^{d} p_\lambda\left(\|\boldsymbol{\beta}_j\|\right), \tag{2.6}$$

where p_λ is a penalty function. An example is the SCAD, which is defined on $\mathbb{R}^+$ by its derivative $p'_\lambda(u) = \lambda I(u \le \lambda) + \frac{(a\lambda - u)_+}{a-1} I(u > \lambda)$ for some $a > 2$. The use of the group SCAD penalty was studied by Wang *et al.* (2008) and Noh & Park (2010) in the setting of spline sieve estimation for longitudinal varying coefficient models. A difficulty arises when p_λ is nonconvex. In such a case, the objective function (2.6) often has multiple local minima. Inspired by Zou & Li (2008) and Noh & Park (2010), one may replace $p_\lambda(\|\boldsymbol{\beta}_j\|)$ in (2.6) by its one-step approximation $p'_\lambda(\|\hat{\boldsymbol{\beta}}_j^{(0)}\|)\|\boldsymbol{\beta}_j\|$, where $\hat{\boldsymbol{\beta}}_j^{(0)}$ are suitable initial estimators of $\boldsymbol{\beta}_j$. The latter approach may enjoy the oracle property for a wide class of penalty functions. Li & Liang (2008) and Kai *et al.* (2011) studied the variable selection problem for the parametric part of the partially linear varying coefficient model (2.4).

Variable selection plays a crucial role in modelling for high-dimensional data whose dimension diverges as the sample size increases. The problem has not been well addressed in nonparametric kernel estimation, especially for the varying coefficient model (1.1) and its variants. The only work that we are aware of is the work of Lam & Fan (2008). In the latter work, they applied a profiling approach to the generalised partially linear varying coefficient model, $g(m(\mathbf{x}, \mathbf{u}, z)) = \mathbf{f}(z)^\top \mathbf{x} + \boldsymbol{\beta}^\top \mathbf{u}$, where g is a link function. However, they allowed only the dimension of $\mathbf{U}^i$ to diverge. Thus, much work needs to be performed for varying coefficient models with high-dimensional data.

2.3 Two-Step Estimation

In this section, we elaborate the idea of two-step estimation introduced by Fan & Zhang (1999) for model (1.1) and studied further by Cai (2002) for model (2.3). Here, we focus on model (1.1). To describe the method, suppose that one obtains $\hat{f}_j$ using a single bandwidth h for all $1 \leq j \leq d$, from the procedure described in Section 2.1. We call these first-step estimators. In the second step, one uses different bandwidths h_j for different f_j. To estimate f_j for a particular j, one refits the residuals $Y^i - \sum_{k \neq j}^d X_k^i \hat{f}_k(Z^i)$ using the bandwidth h_j. In Fan & Zhang (1999), when f_j is estimated for a particular j, local cubic smoothing is applied to f_j and local linear to the others in the first-step estimation, and then local cubic is employed again to fit the residuals $Y^i - \sum_{k \neq j}^d X_k^i \hat{f}_k(Z^i)$ in the second step. In the following, we apply the same order of local polynomial smoothing to all functions in both the first-step and second-step procedures. We treat local constant and local linear smoothing. Extension to higher orders is immediate but needs more involved notation.

Let $\hat{f}_j$ for $1 \leq j \leq d$ be the first-step estimators obtained by Nadaraya–Watson smoothing. For a particular component f_j, the second-step Nadaraya–Watson estimator $\tilde{f}_j(z)$ minimises

$$\sum_{i=1}^{n} \left[Y^i - \sum_{k \neq j}^{d} \hat{f}_k(Z^i) X_k^i - a_j X_j^i \right]^2 K_{h_j}\left(z, Z^i\right)$$

with respect to a_j, where h_j is a bandwidth that is allowed to be different from h in the first-step estimation. For simplicity, we use the same kernel function. In fact, Fan & Zhang (1999) considered the case where $h/h_j \to 0$ as $n \to \infty$. Here, we allow h and h_j to have the same magnitude, thus assume that $h/h_j \to \rho_j$ for some $\rho_j \geq 0$.

To state the asymptotic properties of the two-step estimators, let $N^{jk}(z)$ denote the (j,k)-th entry of $\mathbf{N}(z)^{-1}$. Also, define

$$K_j^*(u) = \int K(t) K(u - \rho_j t)\, dt.$$

Note that if $\rho_j = 0$, then $K_j^* = K$. Denote by $b_k(z)$ the k-th entry of the bias vector $\mathbf{b}(z)$ for the Nasdaraya–Watson estimator introduced in Section 2.1.

Theorem 1. *Let z be a fixed point in $(0,1)$. Assume the conditions (A1), (A3)–(A5) in the Appendix with $q = 1$. Also, assume that $\sigma^2(z)$ is continuous in z and $nh_j^3 \to \infty$ as $n \to \infty$. Then, for the Nadaraya–Watson type $\tilde{f}_j$, it holds that*

$$\begin{aligned} \text{bias}\left(\tilde{f}_j(z) | \mathbb{X}, \mathbb{Z}\right) = \frac{1}{2} h_j^2 \left(\int u^2 K\right) \Bigg[& f_j''(z) + 2\frac{N_{jj}'(z)}{N_{jj}(z)} f_j'(z) + 2\frac{p_0'(z)}{p_0(z)} f_j'(z) \\ & - \rho_j^2 \sum_{k \neq j}^{d} \frac{N_{jk}(z)}{N_{jj}(z)} b_k(z) \Bigg] + o_p\left(h_j^2\right), \end{aligned}$$

$$\operatorname{Var}\left(\tilde{f}_j(z)|\mathbb{X},\mathbb{Z}\right) = n^{-1}h_j^{-1}\frac{\sigma^2(z)}{p_0(z)}\left[N_{jj}(z)^{-1}\left(\int K^2 - \int \left(K_j^*\right)^2\right) + N^{jj}(z)\int \left(K^*\right)^2\right] + o_p\left(n^{-1/2}h_j^{-1/2}\right).$$

Furthermore, $\tilde{f}_j(z)$ has asymptotically a normal distribution.

When $\rho_j = 0$, that is, $h/h_j \to 0$ as $n \to 0$, the leading bias of $\tilde{f}_j$ is equal to that of the oracle estimator, which uses the knowledge of all other functions f_k for $k \neq j$. Also, the leading variance term is simplified to $n^{-1}h_j^{-1}\sigma^2(z)p_0(z)^{-1}N^{jj}(z)\int K^2$. Because $N^{jj}(z) \neq N_{jj}(z)^{-1}$ in general and the leading variance of the oracle estimator is equal to $n^{-1}h_j^{-1}\sigma^2(z)p_0(z)^{-1}N_{jj}(z)^{-1}\int K^2$, the estimator $\tilde{f}_j$ does not have an oracle variance. Note that the variance converges to the oracle variance if $\rho_j \to \infty$ because in this case, $\int \left(K_j^*\right)^2 \to 0$. But then the bias term explodes. Also, because $N^{jj}(z) \geq N_{jj}(z)^{-1}$ and $\int K^2 \geq \int (K_j^*)^2$, the theorem tells that the variance of $\tilde{f}_j(z)$ with $\rho_j > 0$ is smaller than the one with $\rho_j = 0$. For the bias of $\tilde{f}_j(z)$, one cannot tell which case gives a smaller bias, however. The choice $\rho_j = 0$ has an advantage that the bias of $\tilde{f}_j$ does not depend on the shapes of the other nonparametric components $f_k : k \neq j$.

The local linear version of $\tilde{f}_j$ is explicitly given by

$$\tilde{f}_j(z) = \frac{n^{-1}\sum_{i=1}^n \left[S_{2jj}(z) - S_{1j}\left(Z^i - z\right)X_j^i\right]\left[Y^i - \sum_{k\neq j}^d \hat{f}_k\left(Z^i\right)X_k^i\right]K_{h_j}\left(z, Z^i\right)}{S_{0j}(z)S_{2jj}(z) - S_{1j}(z)S_{1jj}},$$

where $S_{lj}(z) = n^{-1}\sum_{i=1}^n \left(Z^i - z\right)^l X_j^i K_{h_j}\left(z, Z^i\right)$ for $l = 0, 1$ and $S_{ljj}(z) = n^{-1}\sum_{i=1}^n \left(Z^i - z\right)^l \left(X_j^i\right)^2 K_{h_j}\left(z, Z^i\right)$ for $l = 1, 2$. For the local linear estimator, we have the following theorem.

Theorem 2. *Let z be a fixed point in* $(0, 1)$. *Assume the conditions in Theorem 1. Then, for the local linear type $\tilde{f}_j$, it holds that*

$$\operatorname{bias}\left(\tilde{f}_j(z)|\mathbb{X},\mathbb{Z}\right) = \frac{1}{2}h_j^2\left(\int u^2 K\right)\left[f_j''(z) - \rho_j^2\sum_{k\neq j}^d \frac{N_{jk}(z)}{N_{jj}(z)}f_k''(z)\right] + o_p\left(h_j^2\right),$$

and the asymptotic conditional variance of $\tilde{f}_j(z)$ admits the same first-order expansion as that of the Nadaraya–Watson type two-step estimator. Furthermore, $\tilde{f}_j(z)$ has an asymptotically normal distribution.

The bandwidth selections ideas that we discussed in Section 2.2 may be implemented for the second-step estimators as well. The two-step approach may be also combined with the idea of shrinkage estimation for variable selection. For this, one may minimise the penalised least squares criteria, (2.5) or (2.6), to obtain the first-step estimators. This would eliminate all insignificant variables X_j for $j \in \mathcal{I}_0$ with probability tending to one, where $\mathcal{I}_0 = \{1 \leq j \leq d : f_j \equiv 0\}$. In the second-step, one includes in the model only those X_j that remains and refit each function in the model by the procedure described earlier. For these two-step estimators, Theorems 1 and 2 remain to hold with $\sum_{k\neq j}^{d} N_{jk}(z)b_k(z)/N_{jj}(z)$ and $\sum_{k\neq j}^{d} N_{jk}(z)f_k''(z)/N_{jj}(z)$, respectively, being replaced by $\sum_{k\neq j,\notin\mathcal{I}_0}^{d} N_{jk}(z)b_k(z)/N_{jj}(z)$ and $\sum_{k\neq j,\notin\mathcal{I}_0}^{d} N_{jk}(z)f_k''(z)/N_{jj}(z)$.

3 Kernel Estimation: Multiple Smoothing Variables

3.1 Models and Methods

There have been a few works on the varying coefficient model (1.2) where different coefficient functions f_j are defined on the different spaces of different variables Z_j. In terms of modelling for real datasets, this is more flexible than model (1.1). Fitting model (1.2) is completely different from fitting the one at (1.1). The reason is that the standard multivariate kernel smoothing, applied locally to each point $\mathbf{z} = (z_1, \ldots, z_d)$ of the covariate vector $\mathbf{Z} = (Z_1, \ldots, Z_d)$, loses the structure of model (1.2) and thus gives multivariate functions of the whole vector $\mathbf{z}$. Because of this, several projection methods have been proposed to obtain proper estimators of the univariate functions f_j.

There have been two approaches to estimate the coefficient functions in (1.2). One is the marginal integration technique of Linton & Nielsen (1995). To estimate f_j, it simply integrates a full-dimensional estimator on the spaces of all smoothing variables except Z_j. In the case of Nadaraya–Watson smoothing, one first minimises

$$n^{-1}\sum_{i=1}^{n}\left(Y^i - \sum_{j=1}^{d} a_j X_j^i\right)^2 K_{h_1}\left(z_1, Z_1^i\right) \times \cdots \times K_{h_d}\left(z_d, Z_d^i\right)$$

for each $\mathbf{z}$. This gives $\hat{a}_j(\mathbf{z})$, which are multivariate functions. The marginal integrations, $\hat{f}_j(z_j) = \int \hat{a}_j(\mathbf{z}) w_j(\mathbf{z}_{-j})\, d\mathbf{z}_{-j}$ give estimators of f_j, where w_j are some weight functions such that $\int w_j\left(\mathbf{z}_{-j}\right) d\mathbf{z}_{-j} = 1$ and $\mathbf{z}_{-j} = \left(z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_d\right)^\top$. However, the technique was found to suffer from the curse of dimensionality because the statistical properties of $\hat{f}_j$ heavily depend on the consistency of $\hat{a}_j$ and thus one requires $nh_1 \times \cdots \times h_d \to \infty$ as $n \to \infty$ for the method to work.

Some earlier works on the marginal integration technique for model (1.2) include Yang *et al.* (2006), Zhang & Li (2007), and Feng *et al.* (2012). Xue & Yang (2006a) studied the marginal integration method for a slightly generalised version of model (1.2), where each $f_j(z_j)$ is replaced by a multivariate function $f_j(\mathbf{z}) = f_{j1}(z_1) + \cdots + f_{jq}(z_q)$ with additive structure,

$$E\left(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}\right) = x_1 \sum_{k=1}^{q} f_{1k}(z_k) + \cdots + x_d \sum_{k=1}^{q} f_{dk}(z_k). \tag{3.1}$$

In this way, one allows the model to include all interaction terms $X_j f_{jk}(Z_k)$ for $1 \le j \le d, 1 \le k \le q$. Some works on testing problems arising from model (1.2) include Yang *et al.* (2006), Ip *et al.* (2007), and Park *et al.* (2011).

The other technique, termed as 'smooth backfitting', was introduced by Mammen *et al.* (1999) for additive regression models. This technique is known to be free of the curse of dimensionality and have several theoretical and numerical advantages over the marginal integration method. Yu *et al.* (2008) studied the smooth backfitting technique for generalised additive models and Lee *et al.* (2010) for additive quantile regression. For the varying coefficient model (1.2), it amounts to minimising the integrated kernel-weighted sum of squares

$$\int n^{-1} \sum_{i=1}^{n} \left(Y^i - \sum_{j=1}^{d} a_j(z_j) X_j^i \right)^2 K_{h_1}\left(z_1, Z_1^i\right) \times \cdots \times K_{h_d}\left(z_d, Z_d^i\right) \, d\mathbf{z}$$

over the space of the function tuples $\mathcal{H} = \{\mathbf{a} = (a_1, \ldots, a_d) : a_j(\mathbf{z}) = a_j(z_j)\}$. Thus, the minimisation is not performed for each $\mathbf{z}$ as in the case of obtaining the full-dimensional estimator.

Lee *et al.* (2012a) provided complete theory for the aforementioned smooth backfitting method. Roca-Pardinas & Sperlich (2010) added a link function to model (1.2). With a link function g, the model for the mean regression function is given by

$$E(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = g^{-1}\left(x_1 f_1(z_1) + \cdots + x_d f_d(z_d)\right). \tag{3.2}$$

They presented a smooth backfitting algorithm but without theory. Recently, Lee *et al.* (2012b) studied the following fully extended version of model (3.2):

$$g\left(m(\mathbf{x})\right) = x_1 \left(\sum_{k \in I_1} f_{1k}(x_k) \right) + \cdots + x_d \left(\sum_{k \in I_d} f_{dk}(x_k) \right), \tag{3.3}$$

where the index sets I_j are known, may not be disjoint, and each I_j does not include j. Note that in this model, any continuous-type variables X_j in the collection of all covariates are allowed to be a smoothing variable, and also each covariate may interact with any of other covariates. Thus, the model is very flexible. If we take $d = 1$ with $X_1 \equiv 1$ and let the only index set I_1 include all indices of other non-constant covariates, then model (3.3) reduces to the generalised additive model studied by Yu *et al.* (2008). Lee *et al.* (2012b) presented a powerful technique of smooth backfitting and gave a complete theory. They also showed that their smooth backfitting estimators outperformed the sieve spline estimators through a simulation study. The main reason for this was that the spline method has too many parameters to estimate. Lee *et al.* (2013) extended the method and theory for model (3.3) to the case where m is a quantile function, that is, $P(Y \leq m(\mathbf{x})|\mathbf{X} = \mathbf{x}) = \alpha$ for some $0 < \alpha < 1$.

3.2 One-Step Estimation

In this section, we discuss the smooth backfitting method of estimating the coefficient functions f_{jk} in model (3.3) that was developed by Lee *et al.* (2012b). For simplicity of notation and presentation, we consider the following model:

$$\begin{aligned} g\left(m(\mathbf{x}, \mathbf{z})\right) &= x_1 \left(\sum_{k=1}^{q} f_{1k}(z_k) \right) + \cdots + x_d \left(\sum_{k=1}^{q} f_{dk}(z_k) \right) \\ &= \mathbf{x}^\top \mathbf{f}_1(z_1) + \cdots + \mathbf{x}^\top \mathbf{f}_q(z_q), \end{aligned} \tag{3.4}$$

where $\mathbf{x} = (x_1, \ldots, x_d)^\top$ and $\mathbf{f}_j(z_j) = \left(f_{1j}(z_j), \ldots, f_{dj}(z_j)\right)^\top$. As in the case of model (1.1), the kernel smoothing for the functions in $\mathbf{f}_j$ acts on the same space of the variable Z_j. Thus, in the one-step estimation, we have to use the same bandwidth, say h_j^0, for all components in each $\mathbf{f}_j$.

For the identification of the coefficient functions f_{jk} in model (3.4), we put the following constraints:

$$\int \mathbf{f}_k(z_k)\omega_k(z_k)\,dz_k = \mathbf{0}, \quad 1 \le k \le q, \tag{3.5}$$

where ω_k are known weight functions. In general, one can pull out a constant vector from each $\mathbf{f}_k$ so that the resulting function vectors satisfy the constraint. These constant vectors can be estimated faster than $\mathbf{f}_k$. Let $\mathbf{h}^0 = (h_1^0, \ldots, h_q^0)^\top$ be the bandwidth vector and $K_{\mathbf{h}^0}(\mathbf{z}, \mathbf{u})$ be the product kernel defined by $K_{\mathbf{h}^0}(\mathbf{z}, \mathbf{u}) = K_{h_1^0}(z_1, u_1) \times \cdots \times K_{h_q^0}(z_q, u_q)$. We use $K_h(z, u) = c(u, h)h^{-1}K((z - u)/h)$, where $c(u, h)$ is chosen so that $\int_0^1 K_h(z, u)\,dz = 1$ for all $u \in [0, 1]$. Let Q denote a quasi-likelihood function defined by $\partial Q(\mu, y)/\partial\mu = (y - \mu)/V(\mu)$, where V is a function for modelling the conditional variance $\sigma^2(\mathbf{x}, \mathbf{z}) = \mathrm{Var}\,(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$ by $V\,(m(\mathbf{x}, \mathbf{z}))$. The Nadaraya–Watson type smooth backfitting estimators $\hat{f}_{jk}$ maximise the integrated kernel-weighted quasi-likelihood

$$L_Q\,(\boldsymbol{\eta}) \equiv \int n^{-1}\sum_{i=1}^{n} Q\left(g^{-1}\left(\boldsymbol{\eta}_1(z_1)^\top\mathbf{X}^i + \cdots + \boldsymbol{\eta}_q\,(z_q)^\top\,\mathbf{X}^i\right), Y^i\right) K_{\mathbf{h}^0}\left(\mathbf{z}, \mathbf{Z}^i\right)\,d\mathbf{z} \tag{3.6}$$

over tuples of functions $(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_q)$, each $\boldsymbol{\eta}_k$ being a vector of univariate functions that satisfies the constraint (3.5). For the local linear version of the one-step estimation, we replace $\boldsymbol{\eta}_k(z_k)$ by $\boldsymbol{\eta}_k\left(z_k, Z_k^i\right) = \boldsymbol{\eta}_{0k}(z_k) + \left(Z_k^i - z_k\right)\boldsymbol{\eta}_{1k}(z_k)$ in L_Q. Then, the local linear smooth backfitting estimators of $\mathbf{f}_k$ and its derivative vectors $\mathbf{f}_k'$ are those $\boldsymbol{\eta}_{0k}$ and $\boldsymbol{\eta}_{1k}$ that maximise the modified L_Q.

The algorithm and the theory for both the Nadaraya–Watson and the local linear one-step estimators may be obtained by adapting the work of Lee *et al.* (2012b) to model (3.4). Indeed, under the conditions (A1)–(A5) in the Appendix, $n^{2/5}\left(\hat{\mathbf{f}}_k(z_k) - \mathbf{f}_k(z_k)\right)$ are jointly asymptotically normal with mean $(\boldsymbol{\beta}_1^{\mathrm{os}}(z_1)^\top, \ldots, \boldsymbol{\beta}_q^{\mathrm{os}}(z_q)^\top)^\top$ and variance $\mathrm{diag}\left(\boldsymbol{\Sigma}_k^{\mathrm{os}}(z_k)\right)$ for some vectors of univariate functions $\boldsymbol{\beta}_k^{\mathrm{os}}$ and some matrices of univariate functions $\boldsymbol{\Sigma}_k^{\mathrm{os}}$. For each $1 \le k \le q$, the matrix $\boldsymbol{\Sigma}_k^{\mathrm{os}}$ for the Nadaraya–Watson estimator is the same as that for the local linear estimator. The bias vectors $\boldsymbol{\beta}_k^{\mathrm{os}}$ for the two estimators are different, however. The local linear estimators have a simpler form. The j-th entry of $\boldsymbol{\beta}_k^{\mathrm{os}}$ for the local linear estimator is given by $\beta_{k:j}^{\mathrm{os}}(z_k) = \left(c_k^0\right)^2\left[f_{jk}''(z_k) - \int f_{jk}''(u)w_k(u)\,du\right]\int u^2K/2$, where $c_k^0 = h_k^0 n^{1/5}$.

3.3 Two-Step Estimation

As in Section 2, we discuss here the two-step estimation of the coefficient functions f_{jk} that uses different bandwidths for different functions. The materials in this section are new and have not been discussed elsewhere. We first treat the Nadaraya–Watson type estimation. We use the one-step estimators discussed in Section 3.2 as first-step estimators. Given those first-step estimators $\left\{\hat{f}_{jk} : 1 \le k \le q,\ 1 \le j \le d\right\}$, the second-step estimator of a specific coefficient function $f_{jk}(z_k)$, which we denote by $\tilde{f}_{jk}(z_k)$, is given by the maximiser of the locally kernel-weighted quasi-likelihood

$$n^{-1}\sum_{i=1}^{n} I_k^i\, Q\left(g^{-1}\left(\sum_{(j',k')\ne(j,k)} \hat{f}_{j'k'}\left(Z_{k'}^i\right)X_{j'}^i + \eta_{jk}X_j^i\right), Y^i\right) K_{h_{jk}}\left(z_k, Z_k^i\right), \tag{3.7}$$

where $I_k^i = I\left(h_{k'}^0 \le Z_{k'}^i \le 1 - h_{k'}^0 \text{ for all } k' \ne k\right)$, and h_{jk} is a bandwidth that is allowed to be different from h_k^0 and also from $h_{j'k'}$ for other $(j', k') \ne (j, k)$. The indicator weights I_k^i are included in (3.7) to exclude the boundary regions where $\hat{f}_{j'k'}$ for $1 \le j' \le d$, $1 \le k' \ne k \le q$ have bias terms of order $n^{-1/5}$. One does not need these indicator weights for local linear kernel smoothing that we discuss later.

For $1 \le k \ne k' \le q$, define

$$\begin{aligned} \mathbf{N}_k(z_k) &= E\left[\frac{\mathbf{X}\mathbf{X}^\top}{V(m(\mathbf{X},\mathbf{Z}))g'(m(\mathbf{X},\mathbf{Z}))^2} \,\middle|\, Z_k = z_k \right], \\ \mathbf{N}_{kk'}(z_k, z_{k'}) &= E\left[\frac{\mathbf{X}\mathbf{X}^\top}{V(m(\mathbf{X},\mathbf{Z}))g'(m(\mathbf{X},\mathbf{Z}))^2} \,\middle|\, Z_k = z_k, Z_{k'} = z_{k'} \right], \\ \mathbf{V}_k(z_k) &= E\left[\frac{\sigma^2(\mathbf{X},\mathbf{Z})}{V(m(\mathbf{X},\mathbf{Z}))^2 g'(m(\mathbf{X},\mathbf{Z}))^2} \mathbf{X}\mathbf{X}^\top \,\middle|\, Z_k = z_k \right]. \end{aligned}$$

Let $N_{k:jj'}$, $N_{kk':jj'}$ and $V_{k:jj'}$ denote the (j, j')-th entry of the matrices $\mathbf{N}_k$, $\mathbf{N}_{kk'}$ and $\mathbf{V}_k$, respectively. Also, let N_k^{jl} denote the (j, l)-th entry of $\mathbf{N}_k^{-1}$. Define

$$K_{jk}^{\#}(u) = \int K(t) K\left(u - \frac{c_k^0}{c_{jk}} t \right) dt.$$

For simplicity, we assume that all h_k^0 and h_{jk} are asymptotic to $n^{-1/5}$. Without loss of generality, we take $h_{jk} = c_{jk} n^{-1/5}$ for some positive constants c_{jk}.

In the next theorem, we give the asymptotic bias and variance of $\tilde{f}_{jk}(z_k)$. To state the theorem, let $\beta_{k:j}^{\mathrm{os}}$ denote the j-th entry of $\boldsymbol{\beta}_k^{\mathrm{os}}$, the asymptotic bias of the one-step Nadaraya–Watson estimator $\hat{\mathbf{f}}_k$ discussed in Section 3.2. Define

$$\begin{aligned} \beta_{k:j}^{\mathrm{ts}}(z_k) = \frac{1}{2}\left(\int u^2 K \right) \Bigg[& c_{jk}^2 \left(f_{jk}''(z_k) + 2\frac{N_{k:jj}'(z_k)}{N_{k:jj}(z_k)} f_{jk}'(z_k) + 2\frac{p_k'(z_k)}{p_k(z_k)} f_{jk}'(z_k) \right) \\ & - \sum_{k'=1,\ne k}^{q} \sum_{j'=1}^{d} \int \frac{N_{kk':jj'}(z_k, z_{k'})}{N_{k:jj}(z_k)} \beta_{k':j'}^{\mathrm{os}}(z_{k'}) p_{k'|k}(z_{k'}|z_k)\, dz_{k'} \\ & - \sum_{j'=1,\ne j} \frac{N_{k:jj'}(z_k)}{N_{k:jj}(z_k)} \beta_{k:j'}^{\mathrm{os}}(z_k) \Bigg], \end{aligned}$$

where $p_{k'|k}(\cdot|z_k)$ is the conditional density of $Z_{k'}$ given $Z_k = z_k$, and

$$\begin{aligned} \Sigma_{k:j}^{\mathrm{ts}}(z_k) = c_{jk}^{-1} p_k(z_k)^{-1} \Bigg[& \left(\int K^2 \right) \frac{V_{k:jj}(z_k)}{N_{k:jj}(z_k)^2} \\ & - 2\left(\int K \cdot K_{jk}^{\#} \right) \left\{ \frac{V_{k:jj}(z_k)}{N_{k:jj}(z_k)^2} - \sum_{l=1}^{d} N_k^{jl}(z_k) \frac{V_{k:lj}(z_k)}{N_{k:jj}(z_k)} \right\} \\ & + \left(\int K_{jk}^{\#2} \right) \Bigg\{ \frac{V_{k:jj}(z_k)}{N_{k:jj}(z_k)^2} - 2\sum_{l=1}^{d} N_k^{jl}(z_k) \frac{V_{k:lj}(z_k)}{N_{k:jj}(z_k)} \\ & + \sum_{l=1}^{d} \sum_{l'=1}^{d} N_k^{jl}(z_k) V_{k:ll'}(z_k) N_k^{l'j}(z_k) \Bigg\} \Bigg]. \end{aligned}$$

If we correctly specify the conditional variance of Y given $\mathbf{X} = \mathbf{x}$ and $\mathbf{Z} = \mathbf{z}$, then the variance term is simplified to

$$\Sigma^{\mathrm{ts}}_{k:j}(z_k) = c_{jk}^{-1} p_k(z_k)^{-1} \left[N_{k:jj}(z_k)^{-1} \left(\int K^2 - \int K^{\#2}_{jk} \right) + N_k^{jj}(z_k) \int K^{\#2}_{jk} \right].$$

Theorem 3. *Let z_k be a fixed point in $(0, 1)$. Assume conditions (A1)–(A5) in the Appendix. Then, for the Nadaraya–Watson type $\tilde{f}_{jk}$, it follows that*

$$n^{2/5} \left(\tilde{f}_{jk}(z_k) - f_{jk}(z_k) \right) \xrightarrow{d} N \left(\beta^{\mathrm{ts}}_{k:j}(z_k), \Sigma^{\mathrm{ts}}_{k:j}(z_k) \right).$$

Now, we discuss the local linear version of the two-step estimators. We maximise the locally kernel-weighted quasi-likelihood at (3.7) with η_{jk} in the summation being replaced by $\eta_{0jk} + \eta_{1jk}(Z_k^i - z_k)$, where $\hat{f}_{j'k'}$ are now the one-step local linear estimators discussed in Section 3.2. This gives the estimators of f_{jk} and their derivatives. To demonstrate the asymptotic distributions of the estimators, we redefine

$$\beta^{\mathrm{ts}}_{k:j}(z_k) = \frac{1}{2} \left(\int u^2 K \right) \left[c_{jk}^2 f''_{jk}(z_k) - \sum_{k'=1, \neq k}^{q} \sum_{j'=1}^{d} \int \frac{N_{kk':jj'}(z_k, z_{k'})}{N_{k:jj}(z_k)} \right.$$
$$\left. \times \beta^{\mathrm{os},*}_{k':j'}(z_{k'}) p_{k'|k}(\cdot|z_k)\, dz_{k'} - \sum_{j'=1, \neq j} \frac{N_{k:jj'}(z_k)}{N_{k:jj}(z_k)} \beta^{\mathrm{os},*}_{k:j'}(z_k) \right],$$

where $\beta^{\mathrm{os},*}_{k:j}(z_k) = \left(c_k^0\right)^2 \left[f''_{jk}(z_k) - \int f''_{jk}(u) w_k(u)\, du \right] \int u^2 K/2$.

Theorem 4. *Let z_k be a fixed point in $(0, 1)$. Assume conditions (A1)–(A5) in the Appendix. Then, for the local linear type $\tilde{f}_{jk}$, it follows that*

$$n^{2/5} \left(\tilde{f}_{jk}(z_k) - f_{jk}(z_k) \right) \xrightarrow{d} N \left(\beta^{\mathrm{ts}}_{k:j}(z_k), \Sigma^{\mathrm{ts}}_{k:j}(z_k) \right).$$

The bias and variance formulas in Theorems 3 and 4 can be used to derive an asymptotically optimal bandwidth h_{jk} for estimating each coefficient function f_{jk}. The formula for the optimal bandwidth includes several unknown quantities that depend on the density p and component functions f_{jk}. These quantities can be replaced by appropriate estimators to obtain a data-driven bandwidth selector. The asymptotic normality results can be used to construct confidence bands for the coefficient functions. It may also justify the use of the bootstrap distributions of the estimators to construct bootstrap confidence bands for f_{jk}.

3.4 Simulation Results

In the simulation study, we considered a binary response Y taking values 0 and 1 and took the following model for the mean function $m(\mathbf{x}, \mathbf{z})$:

$$g\left(m(\mathbf{x}, \mathbf{z})\right) = f_{01}(z_1) + f_{02}(z_2) + x_1 \left(f_{11}(z_1) + f_{12}(z_2)\right) + x_2 \left(f_{21}(z_1) + f_{22}(z_2)\right), \quad (3.8)$$

where $g(u) = \log(u/(1-u))$ is the logit link and $f_{01}(z) = z^2$, $f_{02}(z) = 4(z-0.5)^2$, $f_{11}(z) = z$, $f_{12}(z) = \cos(2\pi z)$, $f_{21}(z) = 1 + e^{2z-1}$, $f_{22}(z) = \sin(2\pi z)$. The covariate X_1 was a

Table 1. *Integrated mean squared errors (IMSE), integrated squared biases (ISB) and integrated variance (IV) of the one-step and two-step estimators for the model (3.8).*

			f_{01}	f_{11}	f_{21}	f_{02}	f_{12}	f_{22}
One-step	$n = 500$	IMSE	0.0293	0.0512	0.0644	0.0532	0.1323	0.1105
		ISB	0.0059	0.0050	0.0337	0.0089	0.0447	0.0493
		IV	0.0234	0.0462	0.0307	0.0443	0.0875	0.0612
	$n = 1,000$	IMSE	0.0177	0.0306	0.0459	0.0315	0.0855	0.0674
		ISB	0.0039	0.0028	0.0267	0.0070	0.0293	0.0340
		IV	0.0138	0.0278	0.0192	0.0245	0.0562	0.0334
Two-step	$n = 500$	IMSE	0.0256	0.0297	0.0731	0.0368	0.0991	0.1102
		ISB	0.0145	0.0125	0.0439	0.0064	0.0062	0.0187
		IV	0.0111	0.0172	0.0291	0.0304	0.0929	0.0915
	$n = 1,000$	IMSE	0.0170	0.0190	0.0533	0.0226	0.0661	0.0625
		ISB	0.0099	0.0071	0.0347	0.0053	0.0057	0.0126
		IV	0.0071	0.0119	0.0186	0.0173	0.0604	0.0499

discrete random variable having Bernoulli (0.5) distribution, X_2 was the standard normal random variable and Z_1 and Z_2 were uniform (0,1) random variables. The four covariates were independent. We consider the estimation of the functions f_{jk} on $[0, 1]$. We chose two sample sizes $n = 500$ and 1,000. The number of replications was 500. For the initial estimate, we used $\hat{\mathbf{f}}^{[0]} = \mathbf{0}$. The functions ω_k in the constraint (3.5) were $\omega_k(z) = I_{[0,1]}(z)$. We computed the theoretically optimal bandwidths, which minimise the asymptotic integrated mean squared errors for the one-step and two-step estimators. For example, the optimal bandwidth h_{jk} in the second step of the two-step estimation is given by $c_{jk}n^{-1/5}$, where c_{jk} minimises $\int_0^1 \left[\beta^{ts}_{k:j}(z_k)^2 + \Sigma^{\mathrm{ts}}_{k:j}(z_k)\right] dz_k$. We used these bandwidths in the simulation. The results are provided in Table 1.

The optimal bandwidths in the one-step estimation and those in the second-step of the two-step estimation were quite different for some component functions. This led to significant gains in estimating those functions in the two-step estimation. For example, in the case of estimating f_{11}, the integrated squared bias and the integrated variance of the two-step estimator $\tilde{f}_{11}$ were balanced much better than those of the one-step estimator $\hat{f}_{11}$. Overall, the two-step estimators have better integrated mean squared errors performance than the one-step estimator. In the case of f_{21}, the one-step estimators are better than the two-step estimator. We found that, in this particular case, the two-step estimator suffered more from the boundary effect on the right end of the unit interval where the true function $f_{21}(z) = 1 + e^{2z-1}$ increases rapidly.

4 Kernel Estimation with Longitudinal Data

Varying coefficient models are particularly useful in longitudinal analysis. A typical framework of longitudinal data consists of a time variable T and a pair of a response Y and a covariate vector $\mathbf{X}$ that are observed over time. In longitudinal analysis, one is often interested in finding how the effects of the covariates on the response change as time evolves. A useful model for this purpose is

$$Y(t) = \mathbf{f}(t)^\top \mathbf{X}(t) + \varepsilon(t), \tag{4.1}$$

where $\mathbf{f} = (f_1, \ldots, f_d)^\top$ is a vector of unknown functions and ε is a mean zero stochastic process. Instead of observing the entire covariate and response processes, one typically observes them intermittently at discrete time points which are random and different for different subjects.

Thus, for the i-th subject ($1 \le i \le n$), one observes $\mathbf{X}^{ij} \equiv \mathbf{X}^i(T_j^i)$ and $Y^{ij} \equiv Y^i(T_j^i)$ at random time points T_j^i, $1 \le j \le n_i$. Thus, for this type of longitudinal data, the varying coefficient model (4.1) can be written as

$$Y^{ij} = \mathbf{f}(T_j^i)^\top \mathbf{X}^{ij} + \varepsilon^{ij}, \tag{4.2}$$

where $\varepsilon^{ij} = \varepsilon^i(T_j^i)$. Typical assumptions in the aforementioned model are that the random time points T_j^i are i.i.d. across all (i, j), and that $\mathbf{X}^{ij}$ as well as ε^{ij} are independent across the n subjects but allowed to be dependent within each subject, so are Y^{ij}.

One can apply the standard kernel smoothing to the model (4.2). If one performs the Nadaraya–Watson smoothing, one minimises

$$N^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(Y^{ij} - \sum_{l=1}^{d} a_l X_l^{ij}\right)^2 K_h(t, T_j^i) \tag{4.3}$$

with respect to a_l, where $N = \sum_{i=1}^n n_i$. In this way, one puts equal weights to all observations. If one wants to put equal weights to all subjects, then one minimises

$$n^{-1}\sum_{i=1}^{n} n_i^{-1}\sum_{j=1}^{n_i}\left(Y^{ij} - \sum_{l=1}^{d} a_l X_l^{ij}\right)^2 K_h(t, T_j^i). \tag{4.4}$$

Hoover *et al.* (1998) studied the method that minimises (4.3), and Wu *et al.* (1998) considered the construction of pointwise and simultaneous confidence regions for f_j.

Wu & Chiang (2000) suggested a componentwise kernel smoothing method in case the covariate $\mathbf{X}$ is time-invariant. With a time-invariant covariate vector, model (4.1) reduces to

$$Y(t) = \mathbf{f}(t)^\top \mathbf{X} + \varepsilon(t). \tag{4.5}$$

To describe the method, note first that $\mathbf{f}(t) = \mathbf{N}^{-1}E(\mathbf{X}Y(t))$, where $\mathbf{N} = E\mathbf{X}\mathbf{X}^\top$. Thus, $f_l(t) = \sum_{k=1}^d N^{lk} E X_k Y(t)$, where N^{lk} is the (l, k)-th entry of $\mathbf{N}^{-1}$. This means that $\sum_{k=1}^d N^{lk} X_k^i Y^{ij}$ for $Y^{ij} = Y^i(T_j^i)$ with T_j^i near the time point t have relevant information about $f_l(t)$. This suggests the minimization of

$$N^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(\sum_{k=1}^{d} \hat{N}^{lk} X_k^i Y^{ij} - \beta_l\right)^2 K_{h_l}(t, T_j^i)$$

to estimate the l-th coefficient function $f_l(t)$ at time t, where $\hat{N}^{lk}$ is the (l, k)-th entry of the inverse of $n^{-1}\sum_{i=1}^n \mathbf{X}^i \mathbf{X}^{i\top}$. The advantage of this approach over the one described in the previous paragraph is that one can use different bandwidths h_l for different functions f_l.

Wu & Yu (2002) gave an overview of nonparametric estimation and inference methods for model (4.2). There have been a few attempts to implement the idea of shrinkage estimation for variable selection in longitudinal varying coefficient models. Among them, Meier & Bühlmann (2007) discussed the LASSO in a simpler setting where the covariate $\mathbf{X}$ is time-invariant and the time points at which one observes the data are fixed and equally spaced. Recently, Daye *et al.* (2012) adapted the fused LASSO idea of Tibshirani *et al.* (2005) for longitudinal varying coefficient models but in the setting of spline sieve estimation.

5 Sieve Estimation and Penalised Likelihood

It has been proposed to use sieve estimators and smoothing splines for varying coefficient models. To represent models (1.1), (1.2) and (3.1) in one form, we write

$$E\left(Y|\mathbf{X},\mathbf{Z}\right) = \sum_{j=1}^{q} \mathbf{X}_j^{\top} \mathbf{f}_j(Z_j),$$

where $\mathbf{X}_j$ has d_j-dimension for some $d_j \geq 1$, and $\mathbf{X}$ is a collection of all covariates contained in $\mathbf{X}_j$ for $1 \leq j \leq q$. Also, it is allowed that $\mathbf{X}_j$ and $\mathbf{X}_k$ for $j \neq k$ have common elements. For example, for model (3.1), $\mathbf{X}_j \equiv \mathbf{X} = (X_1, \ldots, X_d)^{\top}$ for all j. In this framework, sieve estimators are given as minimisers of the least squares criterion

$$n^{-1} \sum_{i=1}^{n} \left[Y_i - \sum_{j=1}^{q} \mathbf{X}_j^{i\top} \mathbf{f}_j(Z_j^i) \right]^2.$$

Here, the minimization runs over finite dimensional classes of functions $\mathbf{f}_1, \ldots, \mathbf{f}_q$, with dimension growing to infinity. Popular choices of the function classes are the spaces of polynomials of degree K or less, the space of trigonometric polynomials of degree K or less and the space of polynomial splines with fixed degree m and K equally spaced knots ('B-spline'). Most work has been performed for the case $q = 1$. These estimators are easy to implement, and a detailed asymptotic theory is available for the understanding of their performance. In this section, our discussion concentrates on the developed asymptotic theory for sieve estimators in varying coefficient models. We will also shortly comment on penalised likelihood estimators.

For sieve estimators, the asymptotic theory contains results on the rates of convergence for several smoothness classes of the functions $\mathbf{f}_1, ..., \mathbf{f}_q$. For a detailed overview on the asymptotic theory of sieve estimation, see Chen (2007). The theory is complete if the bias terms are negligible compared with the variances of the estimators. If both terms are balanced, things become more complicated. For the classical nonparametric regression model $E(Y|X) = f(X)$, this was discussed in Zhou, Shen & Wolfe (1998) and Huang (2003). For B-spline sieve estimators, they gave upper bounds on the bias terms and also provided expansions for the biases under some regularity assumptions. In particular, their assumptions ask that the knot points are nearly equidistant and that the order of the spline basis is related to the degree of smoothness of the underlying regression function f. The expansions have not been available for more complex models such as varying coefficient models.

Some asymptotic properties of sieve estimators in varying coefficient models were studied by Huang *et al.* (2002) and Huang *et al.* (2004). In these papers the rates of convergence were given for B-spline sieve estimators in the longitudinal varying coefficient model (4.2). Furthermore, local asymptotic normality was established under the condition that the number of knot points was chosen so that the bias terms became asymptotically negligible. In Ahmad *et al.* (2005), the latter results were generalised to the partially linear varying coefficient model (2.4). They also showed that the least squares estimator of $\boldsymbol{\beta}$ in (2.4) achieves the semiparametric efficiency bound in the case of homoskedastic errors and proposed a modification of the least squares estimator that is efficient for heteroskedastic errors. Li *et al.* (2011) extended the results on the rate of convergence to the case where the dimension of $\boldsymbol{\beta}$ and the number q of nonparametric components diverges with polynomial rates n^{ρ} for ρ small enough.

The most general varying coefficient model for sieve estimation was treated in Lee *et al.* (2012b). They considered model (3.3). This model also generalises the varying coefficient regression model (3.4) of Xue & Yang (2006a) and Xue & Liang (2010). They used some methods from empirical process theory to show the asymptotic optimal rates for sieve estimators. The basic observation was that entropy conditions on the function classes $\mathcal{F}_{jk}$ for the functions f_{jk} $\left(1 \leq j \leq d; k \in I_j\right)$ carry over to the class of functions for

$$m(\mathbf{x}) = x_1 \left(\sum_{k \in I_1} f_{1k}(x_k) \right) + \cdots + x_d \left(\sum_{k \in I_d} f_{dk}(x_k) \right). \tag{5.6}$$

The argument is similar to the one in the study of additive models. Consider, for example, an additive model with two additive components: $m\left(x_1, x_2\right) = E\left(Y | X_1 = x_1, X_2 = x_2\right) = f_1(x_1) + f_2(x_2)$. Suppose that $f_1 \in \mathcal{F}_1$ and $f_2 \in \mathcal{F}_2$ for some function classes $\mathcal{F}_1$ and $\mathcal{F}_2$. Denote the ε-entropy of the set $\mathcal{F}_j$ by $\log N_j(\varepsilon)$ for $j = 1, 2$, that is, one needs $N_j(\varepsilon)$ balls with radius ε to cover $\mathcal{F}_j$. Then, the ε-entropy of the set $\mathcal{F}_1 \oplus \mathcal{F}_2$ is bounded by $\log N_1(\varepsilon/2) + \log N_2(\varepsilon/2)$. This bound can be used to obtain the rate of convergence for the least squares estimators in the additive model, see van de Geer (2000). In particular, one can compare the rates with those of the least squares estimators in the models $E\left(Y | X_j = x_j\right) = f_j(x_j)$, $j = 1, 2$. One obtains that the rate in the additive model is equal to the slower one of the rates in these two models. Then, one can use the fact that, up to an additive constant, $f_1(x_1)$ is equal to $\int m\left(x_1, x_2\right)\, dx_2$. This shows that the rate of the least squares estimator of the individual additive component f_1 can be bounded by the rate of the estimator of $m = f_1 + f_2$. These arguments apply to varying coefficient models. They can be also used to cover the case of a link function g (generalised varying coefficient models) and the case where another criterion such as a quasi-likelihood is used instead of the least squares, see also the aforementioned discussion on kernel smoothing and the remarks in the succeeding text in this section.

Similar methods can be used to obtain the rates of penalised least squares estimators. We discuss here the penalised least squares estimators with Sobolev penalty, that is, smoothing splines. For model (5.6), the smoothing spline estimator, $\hat{m}(\mathbf{x}) = x_1 \left[\sum_{k \in I_1} \hat{f}_{1k}(x_k)\right] + \cdots + x_d \left[\sum_{k \in I_d} \hat{f}_{dk}(x_k)\right]$ is defined as the minimiser of

$$n^{-1} \sum_{i=1}^{n} \left\{ Y_i - X_1^i \left(\sum_{k \in I_1} f_{1k}\left(X_k^i\right) \right) - \cdots - X_d^i \left(\sum_{k \in I_d} f_{dk}\left(X_k^i\right) \right) \right\}^2 + \lambda_n^2 J(\mathbf{f}),$$

where λ_n^2 is a penalty weight, and $J(\mathbf{f}) = \sum_{k \in I_1} \int D_z^l f_{1k}(z)^2 \, dz + \cdots + \sum_{k \in I_d} \int D_z^l f_{dk}(z)^2 \, dz$ is a penalty term. This minimization problem results in spline functions $\hat{f}_{jk}$ with knots X_{jk}^i, $i = 1, \ldots, n$. Under the assumption that the underlying functions f_{jk} are elements of Sobolev spaces of order l, that is, $\int D_z^l f_{jk}(z)^2 \, dz < \infty$, one obtains $\int (\hat{m}(\mathbf{x}) - m(\mathbf{x}))^2 dx = O_p\left(\lambda_n^2 + n^{-1}\lambda_n^{-1/l}\right)$. This follows directly from the results in van de Geer (2000) based on empirical process methods, see Lee *et al.* (2012b). If the penalty weight λ_n is chosen to have the order $n^{-l/(2l+1)}$, then we obtain $\int (\hat{m}(\mathbf{x}) - m(\mathbf{x}))^2 dx = O_p\left(n^{-2l/(2l+1)}\right)$. By applying the same argument as previously mentioned to the components of m, we obtain $\int \left(\hat{f}_{jk}(z) - f_{jk}(z)\right)^2 dx = O_p\left(n^{-2l/(2l+1)}\right)$. The rate is optimal for functions in Sobolev classes. This result on the rates of smoothing splines also holds for varying coefficient regression models (3.3) with a link function g, see Lee *et al.* (2012b).

There is a limit to the asymptotic analysis of sieve estimators and of smoothing splines based on empirical process theory. It does not give an accurate bound for the rate of an additive component if the order of the entropies of the additive components differ. An important example is the case where an additive component is modelled parametrically. However, in this case, one can proceed with empirical process methods combined with some more complex arguments, see Mammen & van de Geer (1997) and van de Geer (2000) for the related discussion. For an overview on the asymptotic theory of sieve estimation in semiparametric models, see Chen (2007). But nothing seems to be yet known if the entropy orders of nonparametric components differ. Furthermore, to the best of our knowledge, empirical process methods have not been used to obtain local asymptotic results for the least squares estimators of nonparametric components. For these results, one needs explicit expansions of the least squares estimators as it was performed in Zhou *et al.* (1998) and Huang (2003) for the classical nonparametric regression model.

In all aforementioned papers, it has been assumed that i.i.d. data are observed. Varying coefficient models also have natural applications in time series contexts. Consider, for example, the functional coefficient regression model where the conditional expectation of Y is equal to $m(\mathbf{X}, \mathbf{Z}) = \sum_{j=1}^{d} X_j f_j(\mathbf{Z})$. In an autoregression version of this model, one may allow both $\mathbf{X}$ and $\mathbf{Z}$ to contain lagged values of Y. The functional coefficient autoregression model contains several familiar nonlinear parametric time series models as special cases, and it generalises the nonparametric functional autoregressive model of Chen & Tsay (1993). For a general discussion, see also Cai *et al.* (2000) although the latter focused on kernel estimation. In Huang & Shen (2004), B-spline sieve estimators were considered and it was shown that they achieve optimal rates. The asymptotic theory was developed under the assumption of i.i.d. innovations.

We now discuss some more model extensions. In many applications, quantile regression is a more natural or more informative approach than least squares. Instead of assuming that $m(\mathbf{X}, \mathbf{Z}) = \sum_{j=1}^{q} \mathbf{X}_j^{\top} \mathbf{f}_j(Z_j)$ is the conditional mean of Y, one makes the assumption that it is the conditional α-quantile. A nonparametric sieve estimator is then given by the minimiser of

$$n^{-1} \sum_{i=1}^{n} \rho_\alpha \left(Y_i - \sum_{j=1}^{q} \mathbf{X}_j^{i\top} \mathbf{f}_j \left(Z_j^i \right) \right),$$

where $\rho_\alpha(u) = u(\alpha - I(u < 0))$. Here again, the minimization runs over finite dimensional classes of functions $\mathbf{f}_1, \dots, \mathbf{f}_d$, with dimension growing to infinity. A penalised least squares estimator is given by minimising

$$n^{-1} \sum_{i=1}^{n} \rho_\alpha \left(Y_i - \sum_{j=1}^{q} \mathbf{X}_j^{i\top} \mathbf{f}_j \left(Z_j^i \right) \right) + \lambda_n^2 J(\mathbf{f})$$

with penalty term previously defined. In other extensions, the functional ρ_α is replaced by other M-functions. An important example is the quasi-likelihood that we discussed in Section 3 for kernel smoothing. Sieve estimators for the quantile specification were considered in Kim (2007). In the latter paper, optimal rates were derived for B-spline estimators. Extensions of these results were obtained by Wang *et al.* (2009), where a partially linear varying coefficient model was considered. The latter paper contains some results on the rate of convergence of the estimators of the nonparametric parts as well as the parametric parts. In both papers, asymptotic theory was developed for the test of the hypothesis that some components of the function m are constant. However, as in most papers on sieve estimation, no local asymptotic normal limit result for the nonparametric estimators has been established.

6 Heart Disease Data

The dataset was taken from the coronary risk factor baseline survey, which had been conducted on White males aged 15–64 years in a heart disease high-risk region of the Western Cape, South Africa. The aim of the survey was to identify and establish the intensity of ischaemic heart disease risk factors, see Rossouw *et al.* (1983) for details. We analysed a subset of the data, named 'SAheart', which was available from the R package 'ElemStatLearn'. The SAheart dataset contains 462 observations on 10 variables, which include the presence or absence of coronary heart disease (CHD), family history of CHD (F$H = 1$ if one has family history of CHD and F$H = 0$ otherwise), age(AGE), type-A behaviour that is a measure of psycho-social stress (TA), systolic blood pressure (BP) and low density lipoprotein cholesterol (CL).

We considered several varying coefficient models and partially linear varying coefficient models discussed in Sections 2 and 3. We applied the smooth backfitting technique and a sieve method to estimate the nonparametric components in the models. For the sieve estimation, we used cubic splines with a power basis on a equi-spaced knot sequence. These two methods involve tuning parameters, bandwidths and the numbers of knots. We chose the tuning parameters based on a 10-fold cross-validation estimate of the misclassification rate. For this, we partitioned randomly the original data into 10 parts, $\mathcal{X}_1, \ldots, \mathcal{X}_{10}$. Then, for each j, the data with the j-th part removed, $\mathcal{X}_{-j}$, was used for estimation, whereas the j-th part, $\mathcal{X}_j$, was used for validation. Let $\hat{P}_{-j}(\cdot)$ denote the estimate of the conditional probability of the presence of CHD ($Y = 1$) given the predictors $\mathbf{X} = (\text{FH}, \text{BP}, \text{CL}, \text{AGE}, \text{TA})$, obtained by fitting the underlying model to the data $\mathcal{X}_{-j}$ with a choice of bandwidth or knot sequence. We computed the cross-validation criterion

$$\text{err} = \sum_{j=1}^{10} \sum_{i \in \mathcal{X}_j} \left[Y_i - I\left(\hat{P}_{-j}(\mathbf{X}_i) > 0.5 \right) \right]^2 / \left(10 |\mathcal{X}_j| \right) \tag{6.1}$$

and then selected the tuning parameters that minimise the criterion.

To find an appropriate model, we started with the following generalised partially linear additive model as in Hastie & Tibshirani (1987):

$$\log\left(\frac{P}{1-P}\right) = \alpha_0 + \alpha_1 \text{FH} + f_{01}(\text{BP}) + f_{02}(\text{CL}) + f_{03}(\text{AGE}) + f_{04}(\text{TA}). \tag{6.2}$$

Here, $P \equiv P(\mathbf{X})$ denotes the conditional probability of presence of CHD given the predictors $\mathbf{X}$ and f_{0k} are the component functions satisfying the constraints $\int f_{0k}(z_k) dz_k = 0$ for $1 \le k \le 4$. As was observed in Hastie & Tibshirani (1987), we found that the nonparametric fits $\hat{f}_{03}$ and $\hat{f}_{04}$ appeared to be linear. This led us to consider the reduced model

$$\log\left(\frac{P}{1-P}\right) = \alpha_0 + \alpha_1 \text{FH} + \alpha_2 \text{AGE} + \alpha_3 \text{TA} + f_{01}(\text{BP}) + f_{02}(\text{CL}). \tag{6.3}$$

The reduced model seems to be more appropriate than model (6.2) because it produced smaller misclassification rates in both the smooth backfitting estimation (SBF) and spline estimation (SPL). The value of err defined at (6.1) was reduced from 0.2813 to 0.2792 for the SBF and from 0.2813 to 0.2683 for the SPL.

In this field of study, it has been considered as an important task to evaluate the interactions between risk factors, see Hallqvist *et al.* (1996), Hawe *et al.* (2003) and Talmud (2004), for example. For this reason, we were motivated to consider an extended version of model (6.3),

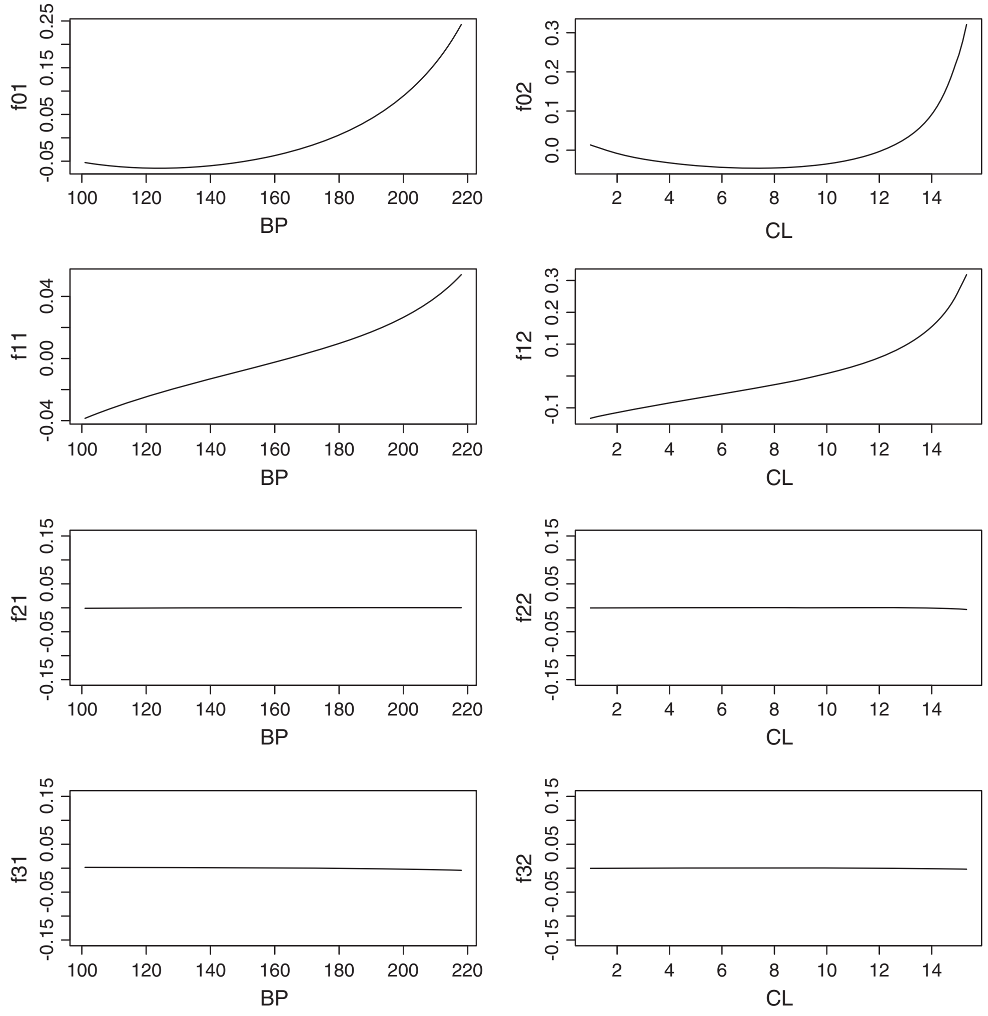

Figure 1. *Estimated coefficient functions* $\hat{f}_{jk}$ *in model (6.4).*

$$\begin{aligned}\log\left(\frac{P}{1-P}\right) = {}& \alpha_0 + f_{01}(\mathrm{BP}) + f_{02}(\mathrm{CL}) + \mathrm{FH} \times [\alpha_1 + f_{11}(\mathrm{BP}) + f_{12}(\mathrm{CL})] \\ & + \mathrm{AGE} \times [\alpha_2 + f_{21}(\mathrm{BP}) + f_{22}(\mathrm{CL})] \\ & + \mathrm{TA} \times [\alpha_3 + f_{31}(\mathrm{BP}) + f_{32}(\mathrm{CL})]\,. \end{aligned} \tag{6.4}$$

This model allows us to see whether the effect of family history, age or stress varies with blood pressure or cholesterol level and includes model (6.3) as a special case. In a separate simulation study, we found that the SPL method gave quite unstable results and were much worse than the SBF method for varying coefficient models, as was also observed in the simulations of Lee *et al.* (2012b). Thus, we do not present the results of the SPL method for this data example.

Figure 1 depicts the estimated functions in model (6.4), and Table 2 contains the estimated values of the coefficients α_j. To assess the statistical significance of the estimated $\hat{\alpha}_j$ and $\hat{f}_{jk}$ in model (6.4), we performed a bootstrap analysis with 100 replications. The 95% bootstrap confidence intervals of α_j are given in Table 2, where we find that all estimated coefficients

Table 2. *Estimates of and 95% bootstrap confidence intervals for α_j in model (6.4).*

	Estimate	Confidence interval
α_0	−5.6353	(−8.9058, −4.0569)
α_1 (FH)	1.0303	(0.6010, 1.7534)
α_2 (Age)	0.0598	(0.0393, 0.0927)
α_3 (TA)	0.0368	(0.0135, 0.0704)

FH, family history; TA, type-A behaviour that is a measure of psycho-social stress.

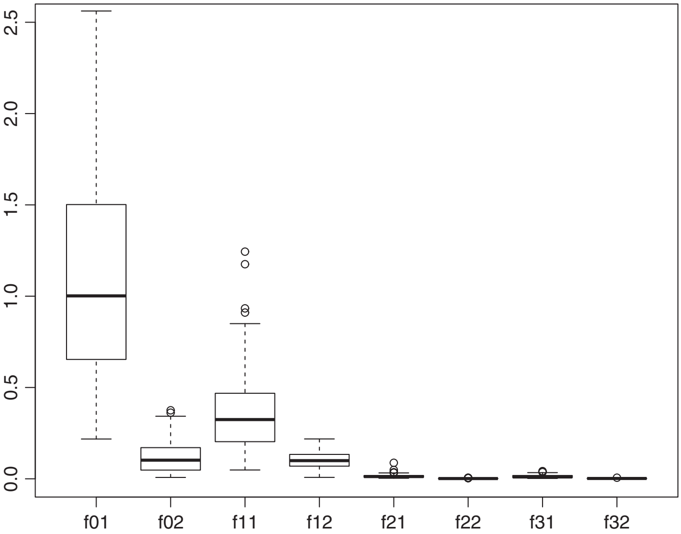

Figure 2. *Boxplot of* $\left[\int\{\hat{f}_{jk}^{(b)}(z)\}^2 dz\right]^{1/2}$, $b = 1, \ldots, 100$, *for model (6.4).*

$\hat{\alpha}_j$ are significant at level 5%. Figure 2 depicts the bootstrap distributions of the L_2-norms of estimated coefficients $\hat{f}_{jk}$. It presents the boxplots of the 100 values of $\|\hat{f}_{jk}^{(b)}\|$, where $\hat{f}_{jk}^{(b)}$ denotes the estimate of f_{jk} from the b-th bootstrap sample. Only the boxplots for $\hat{f}_{01}$, $\hat{f}_{02}$, $\hat{f}_{11}$ and $\hat{f}_{12}$ show a significant departure from zero, which suggested that $f_{21} = f_{22} = f_{31} = f_{32} \equiv 0$ and led us to considering the following reduced model of (6.4):

$$\log\left(\frac{P}{1-P}\right) = \alpha_0 + f_{01}(\mathrm{BP}) + f_{02}(\mathrm{CL}) + \mathrm{FH} \times [\alpha_1 + f_{11}(\mathrm{BP}) + f_{12}(\mathrm{CL})] + \alpha_2 \mathrm{AGE} + \alpha_3 \mathrm{TA}. \tag{6.5}$$

The results of fitting model (6.5) are contained in Table 3 and Figure 3. Table 3 gives the estimates of α_j in model (6.5) and their 95% bootstrap confidence intervals, and Figure 3 depicts the estimated component functions and the 95% pointwise bootstrap confidence bands. Comparing Figures 1 and 3, we find some changes in the shape and smoothness of the estimated functions. This is because the omitted components f_{2k} and f_{3k} affect the estimation of other components and that the bandwidths chosen by the 10-fold cross-validation for a common function in the two models are different. Note also that the vertical scales of the two figures are different.

Because both AGE and TA have significantly positive linear effects in the analysis of model (6.5), people with higher age or stress appear to be more exposed to heart disease. But, their

Table 3. *Estimates of and 95% bootstrap confidence intervals for α_j in model (6.5).*

	Estimate	Confidence interval
α_0	−5.8863	(−8.2698, −4.4229)
α_1 (FH)	1.1896	(0.6134, 1.9224)
α_2 (Age)	0.0641	(0.0499, 0.0862)
α_3 (TA)	0.0376	(0.0128, 0.0620)

FH, family history; TA, type-A behaviour that is a measure of psycho-social stress.

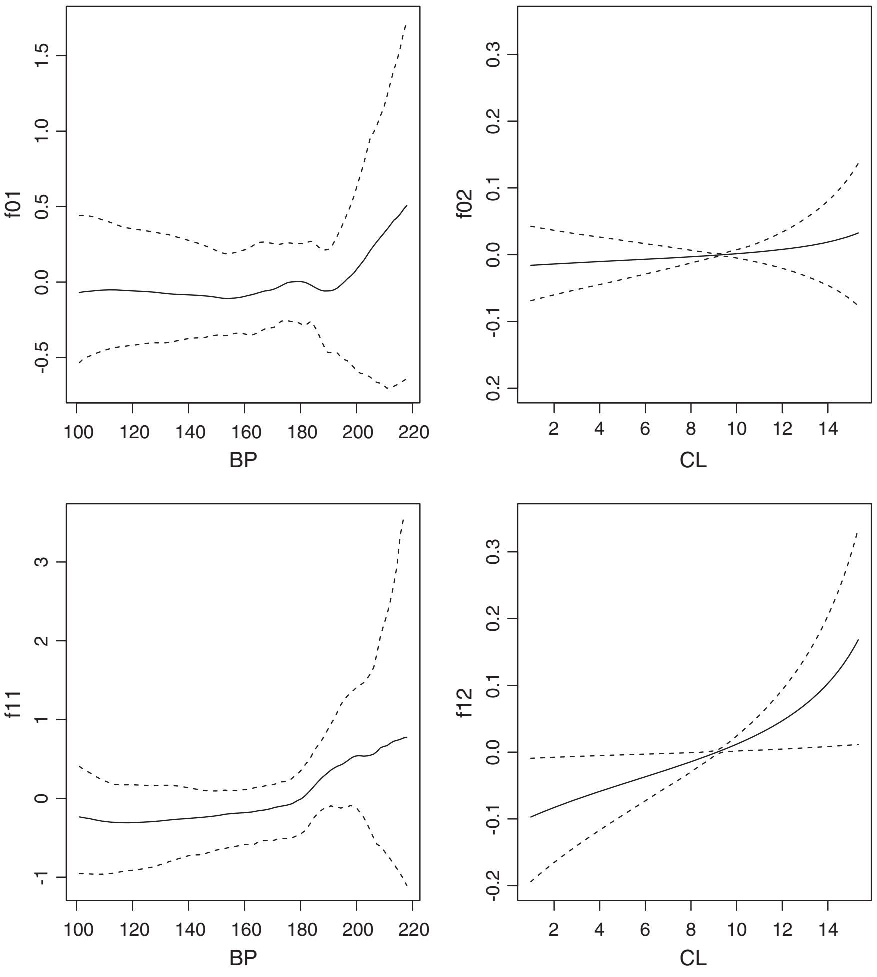

Figure 3. *Estimates of and 95% pointwise bootstrap confidence bands for f_{jk} in model (6.5).*

effects do not change in accordance with blood pressure and cholesterol level because there exists no interaction effect with BP or CL. Regarding the effect of family history, the significantly positive coefficient $\hat{\alpha}_1$ and the increasing trends of $\hat{f}_{11}$(BP) and $\hat{f}_{12}$(CL) suggest that people with positive family history are in more danger of CHD than others and that the effect becomes larger as they have higher blood pressure or cholesterol level. In addition, because the values of $\hat{\alpha}_1 + \hat{f}_{11}(\text{BP}) + \hat{f}_{12}(\text{CL})$ are much larger than the estimated coefficients of AGE and TA, we see that family history is much more influential in increasing the risk of CHD than age and stress. Finally, we find that the estimated coefficient functions of BP, $\hat{f}_{01}$ and $\hat{f}_{11}$, are slightly decreasing when the blood pressure is low, as was also observed in the earlier studies by Hastie & Tibshirani (1987) and Hastie & Tibshirani (1993). This may be partly explained by the retrospectiveness of the data: some of those in this study had been on treatment for reducing their blood pressure after their heart attack, and their measurements were made after the treatment, so that there was a confounded treatment effect.

Acknowledgements

The research of B. U. Park was supported by the NRF grant funded by the Korea government (MEST) with grant no. 2010-0017437. The research of Enno Mammen and Eun Ryung Lee was supported by the Collaborative Research Center SFB 884 'Political Economy of Reforms', financed by the German Science Foundation (DFG), and the research of Young K. Lee was supported by the National Research Foundation of Korea (NRF) with grant no. NRF-2010-616-C00008. AMS 2000 subject classifications: 62G08; 62G20

References

Ahmad, I., Leelahanon, S. & Li, Q. (2005). Efficient estimation of a semiparametric partially linear varying coefficient model. *Ann. Statist.*, **33**, 258–283.

Aitchison, J. & Aitken, C.G.G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, **63**, 413–420.

Breiman, L. & Friedman, J H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.*, **80**, 580–619.

Cai, Z. (2002). Two-step likelihood estimation procedure for varying-coefficient models. *J. Multivariate Anal.*, **82**, 189–209.

Cai, Z., Fan, J. & Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.*, **95**, 888–902.

Cai, Z., Fan, J. & Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *J. Amer. Statist. Assoc.*, **95**, 941–956.

Chen, R. & Tsay, R. (1993). Functional-coefficient autoregressive models. *J. Amer. Statist. Assoc.*, **88**, 298–308.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In *Handbook of Econometrics,* Vol. 6B, Eds. J.J. Heckman & E.E. Leamer, pp. 5549–5631. Amsterdam: North Holland.

Daye, Z.J., Xie, J. & Li, H. (2012). A sparse structured shrinkage estimator for nonparametric varying coefficient model with an application in genomics. *J. Comput. Graph. Statist.*, **21**, 110–133.

Fan, J., Heckman, N.E. & Wand, M.P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.*, **90**, 141–150.

Fan, J. & Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, **11**, 1031–1057.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348–1360.

Fan, J. & Zhang, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.*, **27**, 1491–1518.

Fan, J. & Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scand. J. Stat.*, **27**, 715–731.

Fan, J. & Zhang, W. (2008). Statistical methods with varying coefficient models. *Stat. Interface*, **1**, 179–195.

Feng, J., Huang, Z. & Zhang, R. (2012). Estimation on varying-coefficient partially linear model with different smoothing variables. *Comm. Statist. Theory Methods*, **41**, 516–529.

Galindo, C. D., Liang, H., Kauermann, G. & Carrol, R. J. (2001). Bootstrap confidence intervals for local likelihood, local estimating equations and varying coefficient models. *Statist. Sinica*, **11**, 121–134.

Hallqvist, J., Ahlbom, A., Diderichsen, F. & Reuterwall, C. (1996). How to evaluate interaction between causes: a review of practices in cardiovascular epidemiology. *J. Intern. Med.*, **239**, 377–382.

Härdle, W., Hall, P. & Marron, J. S. (1998). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.*, **83**, 86–101.

Hastie, T. & Tibshirani, R. (1987). Non-parametric logistic and proportional odds regression. *J. R. Stat. Soc. Ser. C.*, **36**, 260–276.

Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models. *J. R. Stat. Soc. Ser. B.*, **55**, 757–796.

Hawe, E., Talmud, P. J., Miller, G. J. & Humphries, S. E. (2003). Family history is a coronary heart disease risk factor in the second northwick park heart study. *Ann. Human Genetics*, **67**, 97–106.

Honda, T. (2004). Quantile regression in varying coefficient models. *J. Statist. Plann. Inference*, **121**, 113–125.

Hoover, D., Rice, J., Wu, C. & Yang, L. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809–822.

Hu, T. & Xia, Y. (2012). Adaptive semi-varying coefficient model selection. *Statist. Sinica*, **22**, 575–599.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *Ann. Statist.*, **31**, 1600–1635.

Huang, J. Z. & Shen, H. (2004). Functional coefficient regression models for non-linear time series: a polynomial spline approach. *Scand. J. Stat.*, **31**, 515–534.

Huang, J. Z., Wu, C. O. & Zhou, L. (2002). Varying coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, **89**, 111–128.

Huang, J. Z., Wu, C. O. & Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica*, **14**, 763–788.

Ip, W., Wong, H. & Zhang, R. (2007). Generalized likelihood ratio test for varying-coefficient models with different smoothing variables. *Comput. Statist. Data Anal.*, **51**, 4543–4561.

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

Kauermann, G. & Tutz, G. (1999). On model diagnostics using varying coefficient models. *Biometrika*, **86**, 119–128.

Kauermann, G. & Tutz, G. (2000). Local likelihood estimation in varying-coefficient models including additive bias correction. *J. Nonparametr. Stat.*, **12**, 343–371.

Kim, M.-O. (2007). Quantile regression with varying coefficients. *Ann. Statist.*, **35**, 92–108.

Lam, C. & Fan, J. (2008). Profile-kernel likelihood inference with diverging number of parameters. *Ann. Statist.*, **36**, 2232–2260.

Lee, Y. K., Mammen, E. & Park, B. U. (2010). Backfitting and smooth backfitting for additive quantile models. *Ann. Statist.*, **38**, 2857–2883.

Lee, Y. K., Mammen, E. & Park, B. U. (2012a). Projection-type estimation for varying coefficient regression models. *Bernoulli*, **18**, 177–205.

Lee, Y. K., Mammen, E. & Park, B. U. (2012b). Flexible generalized varying coefficient regression models. *Ann. Statist.*, **40**, 1906–1933.

Lee, Y. K., Mammen, E. & Park, B. U. (2013). Backfitting and smooth backfitting in varying coefficient quantile regression. *Econometrics J.* DOI: 10.1111/ectj.12017.

Li, R. & Liang, H. (2008). Variable selection in semiparametric regression modeling. *Ann. Statist.*, **36**, 261–286.

Li, G., Xue, L. & Lian, H. (2011). Semi-varying coefficient models with a diverging number of components. *J. Multivariate Anal.*, **102**, 1166–1174.

Linton, O. & Nielsen, J. P. (1995). A kernel method of structured nonparametric regression based on marginal integration. *Biometrika*, **83**, 529–540.

Mammen, E., Linton, O. & Nielsen, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.*, **27**, 1443–1490.

Mammen, E. & Park, B. U. (2005). Bandwidth selection for smooth backfitting in additive models. *Ann. Statist.*, **33**, 1260–1294.

Mammen, E. & van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist.*, **25**, 1014–1035.

Meier, L. & Bühlmann, P. (2007). Smoothing l_1-penalized estimators for high-dimensional time-course data. *Electron. J. Stat.*, **1**, 597–615.

Noh, H. S. & Park, B. U. (2010). Sparse varying coefficient models for longitudinal data. *Statist. Sinica*, **20**, 1183–1202.

Park, B. U., Hwang, J. H. & Park, M. S. (2011). Testing in nonparametric varying coefficient additive models. *Statist. Sinica*, **21**, 749–778.

Roca-Pardinas, J. & Sperlich, S. (2010). Feasible estimation in generalized structured models. *Stat. Comput.*, **20**, 367–379.

Rossouw, J. E., Du Plessis, J. P., Benadé, A. J., Jordaan, P. C., Kotzé, J. P., Jooste, P. L. & Ferreira, J. J. (1983). Coronary risk factor screening in three rural communities. The CORIS baseline study. *S. Afr. Med. J.*, **64**, 430–436.

Talmud, P. J. (2004). How to identify gene-environment interactions in a multifactorial disease: CHD as an example. *Proc. Nutr. Soc.*, **63**, 5–10.

Tang, Q. & Wang, J. (2005). L_1-estimation for varying coefficient models. *Statistics*, **39**, 389–404.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Ser. B.*, **58**, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B.*, **67**, 91–108.

van de Geer, S. (2000). *Empirical Processes in M-Estimation.* Cambridge: Cambridge University Press.

Wang, L., Kai, B. & Li, R. (2009). Local rank inference for varying coefficient models. *J. Amer. Statist. Assoc.*, **104**, 1631–1645.

Wang, L., Li, H. & Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.*, **103**, 1556–1569.

Wang, H. & Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *J. Amer. Statist. Assoc.*, **104**, 747–757.

Wang, H. J., Zhu, Z. & Zhou, J. (2009). Quantile regression in partially linear varying coefficient models. *Ann. Statist.*, **37**, 3841–3866.

Wu, C. & Chiang, C. T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statist. Sinica*, **10**, 433–456.

Wu, C., Chiang, C. T. & Hoover, D. R. (1998). Asymptotic confidence regions for kernel model smoothing of a varying-coefficient with longitudinal data. *J. Amer. Statist. Assoc.*, **93**, 1458–1476.

Wu, C. & Yu, K. (2002). Nonparametric varying-coefficient models for the analysis of longitudinal data. *Int. Stat. Rev.*, **70**, 373–393.

Xue, L. & Liang, H. (2010). Polynomial spline estimation for a generalized additive coefficient model. *Scand. J. Stat.*, **37**, 26–46.

Xue, L. & Yang, L. (2006a). Estimation of semiparametric additive coefficient model. *J. Statist. Plann. Inference*, **136**, 2506–2534.

Yang, L., Park, B. U., Xue, L. & Härdle, W. (2006). Estimation and testing for varying coefficients in additive models with marginal integration. *J. Amer. Statist. Assoc.*, **101**, 1212–1227.

You, J. & Chen, G. (2006). Estimation of a semiparametric varying-coefficient partially linear errors-in-variables model. *J. Multivariate Anal.*, **97**, 324–341.

You, J., Zhou, Y. & Chen, G. (2006). Corrected local polynomial estimation in varying-coefficient models with measurement errors. *Canad. J. Statist.*, **34**, 391–410.

Yu, K., Park, B. U. & Mammen, E. (2008). Smooth backfitting in generalized additive models. *Ann. Statist.*, **36**, 228–260.

Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B.*, **68**, 46–67.

Zhang, W. & Lee, S. (2000). Variable bandwidth selection in varying-coefficient models. *J. Multivariate Anal.*, **74**, 116–134.

Zhang, R. & Li, G. (2007). Averaged estimation of functional-coefficient regression models with different smoothing variables. *Statist. Probab. Lett.*, **77**, 455–461.

Zhang, W. & Peng, H. (2010). Simultaneous confidence band and hypothesis test in generalised varying-coefficient models. *J. Multivariate Anal.*, **101**, 1656–1680.

Zhou, Y. & Liang, H. (2009). Statistical inference for semiparametric varying-coefficient partially linear models with error-prone linear covariates. *Ann. Statist.*, **37**, 427–458.

Zhou, S., Shen, X. & Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *Ann. Statist.*, **26**, 1760–1782.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *J. Amer. Statist. Assoc.*, **101**, 1418–1429.

Zou, H. & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, **36**, 1509–1566.

Appendix A: Technical Details

The proofs of Theorems 1 and 2 are less involved than those of Theorems 3 and 4, respectively, and their main ideas are the same as those of the latter two. Also, the proof of Theorem 4 is similar to that of Theorem 3. Thus, we only give the proof of Theorem 3 here.

We first collect the assumptions we use for the theory in Sections 2 and 3. Define $Q_j(u,y) = \partial^j Q\left(g^{-1}(u),y\right)/\partial u^j$.

(A1) For simplicity, we assume that the support of the variable $\mathbf{Z}$ is equal to $[0,1]^q$ and that the elements of $\mathbf{X}$ are supported on a finite set or a bounded interval. The joint density function $p(\mathbf{x}, \mathbf{z})$ of $(\mathbf{X}, \mathbf{Z})$ is bounded away from zero and infinity on its support and partially continuously differentiable in $\mathbf{z}$ for all $\mathbf{x}$. The smallest eigenvalues of $E(\mathbf{X}\mathbf{X}^\top | Z_k = z_k)$ for $1 \le k \le q$ are bounded away from zero on $[0, 1]$.

(A2) The quasi-likelihood function $Q(\mu, y)$ is three times continuously differentiable with respect to μ for all y in the range of Y and $Q_2(u, y) < 0$ for $u \in \mathbb{R}$ and y in the range of Y. The link function g is three times continuously differentiable, V is twice continuously differentiable, and the conditional variance function $\sigma^2(\mathbf{x}, \mathbf{z})$ is continuous in $\mathbf{z}$ for each $\mathbf{x}$. The function V and the derivative g' are bounded away from zero. The higher-order derivatives g'' and g''' are bounded. The weight functions w_j are continuously differentiable, fulfil $w_j(0) = w_j(1) = 0$, $w_j(x_j) \ge 0$ for $x_j \in [0, 1]$ and $\int w_j(x_j)\, dx_j > 0$ for $1 \le j \le q$.

(A3) The components of $\mathbf{f}_k$ are twice continuously differentiable.

(A4) $E|Y|^\alpha < \infty$ for some $\alpha > 5/2$.

(A5) The kernel function K is bounded, symmetric about zero, has compact support, say $[-1, 1]$, and is Lipschitz continuous.

To prove Theorem 3, define

$$\hat{\mathbf{F}}_k(z_k) = \int n^{-1} \sum_{i=1}^{n} Q_1\left(\mathbf{X}^{i\top} \sum_{l=1}^{q} \mathbf{f}_l(z_l),\, Y^i\right) \mathbf{X}^i K_{\mathbf{h}^0}(\mathbf{z}, \mathbf{Z}^i)\, d\mathbf{z}_{-k},$$

$$\hat{\mathbf{W}}_{kk}(z_k) = -\int n^{-1} \sum_{i=1}^{n} Q_2\left(\mathbf{X}^{i\top} \sum_{l=1}^{q} \mathbf{f}_l(z_l),\, Y^i\right) \mathbf{X}^i \mathbf{X}^{i\top} K_{\mathbf{h}^0}\left(\mathbf{z}, \mathbf{Z}^i\right)\, d\mathbf{z}_{-k}.$$

Let $\tilde{\boldsymbol{\delta}}_k(z_k) = \hat{\mathbf{W}}_{kk}(z_k)^{-1} \hat{\mathbf{F}}_k(z_k)$ and $\tilde{\boldsymbol{\delta}}_k^A(z_k) = \tilde{\boldsymbol{\delta}}_k(z_k) - E\left[\tilde{\boldsymbol{\delta}}_k(z_k) | \mathbf{X}^1, \ldots, \mathbf{X}^n, \mathbf{Z}^1, \ldots, \mathbf{Z}^n\right]$. Let $\tilde{\delta}_{k:j}^A$ denote the j-th entry of the vector $\tilde{\boldsymbol{\delta}}_k^A$. Then, for a fixed $z_k \in (0, 1)$, we obtain

$$\begin{aligned}
\tilde{f}_{jk}(z_k) - f_{jk}(z_k) = {} & \left[N_{k:jj}(z_k) p_k(z_k) + o_p(1)\right]^{-1} \\
& \times \left[n^{-1} \sum_{i=1}^{n} I_k^i\, Q_1\left(\sum_{l=1}^{q} \mathbf{f}_l\left(Z_l^i\right)^\top \mathbf{X}^i, Y^i\right) X_j^i\, K_{h_{jk}}\left(z_k, Z_k^i\right) \right. \\
& \quad + \sum_{(j',k') \ne (j,k)} n^{-1} \sum_{i=1}^{n} I_k^i\, Q_2\left(\sum_{l=1}^{q} \mathbf{f}_l(Z_l^i)^\top \mathbf{X}^i, Y^i\right) \\
& \quad \times \left[\tilde{\delta}_{k':j'}^A\left(Z_{k'}^i\right) + n^{-2/5} \beta_{k':j'}^{\mathrm{os}}\left(Z_{k'}^i\right)\right] X_j^i X_{j'}^i K_{h_{jk}}\left(z_k, Z_k^i\right) \\
& \quad + n^{-1} \sum_{i=1}^{n} I_k^i\, Q_2\left(\sum_{l=1}^{q} \mathbf{f}_l\left(Z_l^i\right)^\top \mathbf{X}^i, Y^i\right) \left\{f_{jk}(z_k) - f_{jk}\left(Z_k^i\right)\right\} \\
& \quad \left. \times \left(X_j^i\right)^2 K_{h_{jk}}\left(z_k, Z_k^i\right) \right] + o_p\left(n^{-2/5}\right).
\end{aligned}$$

The aforementioned expansion may be obtained from an expansion of $\hat{\mathbf{f}}_k$ as in the proof of Theorem 3 of Lee *et al.* (2012b) and by applying the standard technical arguments in kernel smoothing to local quasi-likelihood estimation, see Fan *et al.* (1995), for example.

Now, we approximate each term in the aforementioned expansion of $\tilde{f}_{jk}(z_k) - f_{jk}(z_k)$. First, we obtain

$$
\begin{aligned}
& n^{-1}\sum_{i=1}^{n} I_k^i \, Q_2\left(\sum_{l=1}^{q} \mathbf{f}_l(Z_l^i)^\top \mathbf{X}^i, Y^i\right) \mathbf{X}^i \mathbf{X}^{i\top} \boldsymbol{\beta}_{k'}^{\mathrm{os}}\left(Z_{k'}^i\right) K_{h_{jk}}\left(z_k, Z_k^i\right) \\
&\quad = -\int \mathbf{N}_{kk'}\left(z_k, z_{k'}\right) \boldsymbol{\beta}_{k'}^{\mathrm{os}}(z_{k'}) p_{kk'}\left(z_k, z_{k'}\right)\, dz_{k'} + o_p(1),\ k' \neq k, \\
& n^{-1}\sum_{i=1}^{n} I_k^i \, Q_2\left(\sum_{l=1}^{q} \mathbf{f}_l\left(Z_l^i\right)^\top \mathbf{X}^i, Y^i\right) \mathbf{X}^i \mathbf{X}^{i\top} \boldsymbol{\beta}_{k}^{\mathrm{os}}\left(Z_{k}^i\right) K_{h_{jk}}\left(z_k, Z_k^i\right) \\
&\quad = -\mathbf{N}_k(z_k) \boldsymbol{\beta}_k^{\mathrm{os}}\left(z_k\right) p_k(z_k) + o_p(1).
\end{aligned}
\tag{A.1}
$$

Also, we have

$$
\begin{aligned}
& n^{-1}\sum_{i=1}^{n} I_k^i \, Q_2\left(\sum_{l=1}^{q} \mathbf{f}_l\left(Z_l^i\right)^\top \mathbf{X}^i, Y^i\right) \left\{f_{jk}(z_k) - f_{jk}\left(Z_k^i\right)\right\} \left(X_j^i\right)^2 K_{h_{jk}}\left(z_k, Z_k^i\right) \\
&\quad = \frac{1}{2} h_{jk}^2 \left(\int u^2 K\right) \Big[N_{k:jj}(z_k) f_{jk}''(z_k) p_k(z_k) + 2\, N_{k:jj}'(z_k) f_{jk}'(z_k) p_k(z_k) \\
&\qquad + 2\, N_{k:jj}(z_k) f_{jk}'(z_k) p_k'(z_k)\Big] + o_p\left(n^{-2/5}\right).
\end{aligned}
\tag{A.2}
$$

These two approximations (A.1) and (A.2) give the bias expansion in Theorem 3.

The asymptotic variance of the estimator $\tilde{f}_{jk}(z_k)$ comes from

$$
\begin{aligned}
S_{jk}(z_k) \equiv n^{-1}\sum_{i=1}^{n} I_k^i \Bigg[& Q_1\left(\sum_{l=1}^{q} \mathbf{f}_l(Z_l^i)^\top \mathbf{X}^i, Y^i\right) X_j^i \, K_{h_{jk}}\left(z_k, Z_k^i\right) \\
& + \sum_{(j',k') \neq (j,k)} Q_2\left(\sum_{l=1}^{q} \mathbf{f}_l(Z_l^i)^\top \mathbf{X}^i, Y^i\right) \tilde{\delta}_{k':j'}^{A}\left(Z_{k'}^i\right) X_j^i X_{j'}^i K_{h_{jk}}\left(z_k, Z_k^i\right)\Bigg].
\end{aligned}
$$

Let $\epsilon^i = Y^i - m(\mathbf{X}^i, \mathbf{Z}^i)$. Then,

$$
\begin{aligned}
\tilde{\boldsymbol{\delta}}_k^A(z_k) &= \hat{\mathbf{W}}_{kk}(z_k)^{-1} n^{-1}\sum_{i=1}^{n} \epsilon^i \int \tau\left(m\left(\mathbf{X}^i, \mathbf{z}\right)\right)^{-1} \mathbf{X}^i K_{\mathbf{h}^0}\left(\mathbf{z}, \mathbf{Z}^i\right)\, d\mathbf{z}_{-k} \\
&= \mathbf{N}_k(z_k)^{-1} p_k(z_k)^{-1} n^{-1}\sum_{i=1}^{n} \epsilon^i \tau\left(m\left(\mathbf{X}^i, \mathbf{Z}^i\right)\right)^{-1} \mathbf{X}^i K_{h_k^0}\left(z_k, Z_k^i\right) + o_p\left(n^{-2/5}\right)
\end{aligned}
\tag{A.3}
$$

uniformly for $z_k \in \left[h_k^0, 1 - h_k^0\right]$, where $\tau(u) = V(u) g'(u)$. The aforementioned uniform approximation can be obtained from $\sup_{z_k \in [h_k^0, 1-h_k^0]} \|\hat{\mathbf{W}}_{kk}(z_k) - \mathbf{N}_k(z_k) p_k(z_k)\| = o_p(1)$ and

$$
\begin{aligned}
& \sup_{z_k \in [0,1]} \left| n^{-1}\sum_{i=1}^{n} \epsilon^i \int \left[\tau\left(m\left(\mathbf{X}^i, \mathbf{z}\right)\right)^{-1} - \tau\left(m\left(\mathbf{X}^i, \mathbf{Z}^i\right)\right)^{-1}\right] \mathbf{X}^i K_{\mathbf{h}^0}\left(\mathbf{z}, \mathbf{Z}^i\right)\, d\mathbf{z}_{-k} \right| \\
&\quad = O_p\left(n^{-3/5}\sqrt{\log n}\right),
\end{aligned}
$$

where $\|\cdot\|$ denotes the Hilbert–Schmidt norm. Also, we have

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n} I_k^i Q_1\left(\sum_{l=1}^{q}\mathbf{f}_l\left(Z_l^i\right)^{\top}\mathbf{X}^i, Y^i\right) X_j^i\, K_{h_{jk}}\left(z_k, Z_k^i\right) \\ & \quad = n^{-1}\sum_{i=1}^{n} I_k^i \epsilon^i\, \tau\left(m\left(\mathbf{X}^i,\mathbf{Z}^i\right)\right)^{-1} X_j^i\, K_{h_{jk}}\left(z_k, Z_k^i\right). \end{aligned} \tag{A.4}$$

Now, we define

$$\begin{aligned} J_{jk:j'k'l}(u,z) = n^{-1}\sum_{i=1}^{n}&\left[\frac{X_j^i X_{j'}^i}{V\left(m\left(\mathbf{X}^i,\mathbf{Z}^i\right)\right) g'\left(m\left(\mathbf{X}^i,\mathbf{Z}^i\right)\right)^2}\right]\frac{N_{k'}\left(Z_{k'}^i\right)^{j'l}}{p_{k'}\left(Z_{k'}^i\right)} \\ & \times K_{h_{k'}^0}\left(Z_{k'}^i,u\right) K_{h_{jk}}\left(z, Z_k^i\right), \end{aligned}$$

where $N_k^{j'l}$ denotes the (j',l)-th entry of the matrix $\mathbf{N}_k^{-1}$. From the expressions (A.3) and (A.4), we can write $S_{jk}(z_k) = n^{-1}\sum_{i=1}^{n}\xi_{jk}^i(z_k)\epsilon^i$ with the following definition of ξ_{jk}^i:

$$\xi_{jk}^i(z_k) = \tau\left(m\left(\mathbf{X}^i,\mathbf{Z}^i\right)\right)^{-1}\left[X_j^i\, K_{h_{jk}}\left(z_k, Z_k^i\right) - \sum_{(j',k')\neq(j,k)}\sum_{l=1}^{d} J_{jk:j'k'l}\left(Z_{k'}^i, z_k\right) X_l^i\right],$$

which involves only $\left(\mathbf{X}^i,\mathbf{Z}^i\right)$, not ϵ^i. We observe that the following approximations of $J_{jk:j'k'l}(u,z)$ hold uniformly for $u\in\left[h_{k'}^0, 1-h_{k'}^0\right]$

$$\begin{aligned} J_{jk:j'kl}(u,z) &= N_{k:jj'}(u) N_k^{j'l}(u) K_{jk,h_{jk}}^{\#}(z,u)\left[1+o_p(1)\right], \quad k'=k \\ J_{jk:j'k'l}(u,z) &= N_{kk':jj'}(z,u) N_{k'}^{j'l}(u)\frac{p_{kk'}(z,u)}{p_{k'}(u)}\left[1+o_p(1)\right], \quad k'\neq k, \end{aligned} \tag{A.5}$$

where $K_{jk,h_{jk}}^{\#}(z,u)$ is defined as $K_{h_{jk}}(z,u)$ but with K being replaced by $K_{jk}^{\#}$. The second result of (A.5) implies that the contributions by $J_{jk:j'k'l}$ with $k'\neq k$ in ξ_{jk}^i to the variance of S_{jk} are of smaller order than those by $J_{jk:j'kl}$. Thus, we may neglect these terms in the approximation of $\mathrm{Var}\left[S_{jk}(z_k)|\mathbf{X}^1,\ldots,\mathbf{X}^n,\mathbf{Z}^1,\ldots,\mathbf{Z}^n\right]$. We obtain

$$\begin{aligned} & n^{-1}h_{jk}\sum_{i=1}^{n}\frac{\sigma^2\left(\mathbf{X}^i,\mathbf{Z}^i\right)}{\tau\left(m\left(\mathbf{X}^i,\mathbf{Z}^i\right)\right)^2}\sum_{j'\neq j}^{d}\sum_{l=1}^{d} J_{jk:j'kl}\left(Z_k^i, z_k\right) X_l^i X_j^i K_{h_{jk}}\left(z_k, Z_k^i\right) \\ & \quad = p_k(z_k)\left(\int K\cdot K_{jk}^{\#}\right)\left[E\left(\frac{\sigma^2(\mathbf{X},\mathbf{Z})}{\tau\left(m\left(\mathbf{X},\mathbf{Z}\right)\right)^2}X_j^2\,\middle|\, Z_k = z_k\right)\right. \\ & \qquad \left. - N_{k:jj}(z_k)\sum_{l=1}^{d} N_k^{jl}(z_k) E\left(\frac{\sigma^2(\mathbf{X},\mathbf{Z})}{\tau\left(m\left(\mathbf{X},\mathbf{Z}\right)\right)^2}X_l X_j\,\middle|\, Z_k = z_k\right)\right] + o_p(1). \end{aligned} \tag{A.6}$$

We also obtain

$$\begin{aligned}
&n^{-1} h_{jk} \sum_{i=1}^{n} \frac{\sigma^2\left(\mathbf{X}^i, \mathbf{Z}^i\right)}{\tau\left(m\left(\mathbf{X}^i, \mathbf{Z}^i\right)\right)^2} \left[\sum_{j' \neq j}^{d} \sum_{l=1}^{d} J_{jk:j'kl}\left(Z_k^i, z_k\right) X_l^i \right]^2 \\
&\quad = p_k(z_k) \int \left(K_{jk}^{\#}\right)^2 \left[E\left(\frac{\sigma^2(\mathbf{X}, \mathbf{Z})}{\tau\left(m\left(\mathbf{X}, \mathbf{Z}\right)\right)^2} X_j^2 \,\middle|\, Z_k = z_k \right) \right. \\
&\qquad\qquad - 2\, N_{k:jj}(z_k) \sum_{l=1}^{d} N_k^{jl}(z_k) E\left(\frac{\sigma^2(\mathbf{X}, \mathbf{Z})}{\tau\left(m\left(\mathbf{X}, \mathbf{Z}\right)\right)^2} X_l X_j \,\middle|\, Z_k = z_k \right) \\
&\qquad\qquad \left. + N_{k:jj}(z_k)^2 \sum_{l=1}^{d} \sum_{l'=1}^{d} N_k^{jl}(z_k) E\left(\frac{\sigma^2(\mathbf{X}, \mathbf{Z})}{\tau\left(m\left(\mathbf{X}, \mathbf{Z}\right)\right)^2} X_l X_{l'} \,\middle|\, Z_k = z_k \right) N_k^{l'j}(z_k) \right] \\
&\qquad + o_p(1). \qquad\qquad (A.7)
\end{aligned}$$

The approximations (A.6) and (A.7) give the variance expansion in Theorem 3.

[*Received June 2012, accepted June 2013*]