Published in final edited form as:
Stat Interface. 2008 ; 1(1): 179–195.

Statistical Methods with Varying Coefficient Models

Jianqing Fan* and
Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544

Wenyang Zhang
Department of Mathematical Sciences, University of Bath, UK

Abstract

The varying coefficient models are very important tool to explore the dynamic pattern in many scientific areas, such as economics, finance, politics, epidemiology, medical science, ecology and so on. They are natural extensions of classical parametric models with good interpretability and are becoming more and more popular in data analysis. Thanks to their flexibility and interpretability, in the past ten years, the varying coefficient models have experienced deep and exciting developments on methodological, theoretical and applied sides. This paper gives a selective overview on the major methodological and theoretical developments on the varying coefficient models.

Keywords

Varying coefficient models; local linear modelling; bandwidth selection; cross-validation; confidence band; hypothesis test; semivarying coefficient models; exponential family; generalized varying coefficient models; local maximum likelihood; nonlinear time series; longitudinal data analysis; Cox models; local partial likelihood

1 Why varying coefficient models?

1.1 Theoretical background

Parametric statistical inference always necessitates some model assumptions, linearity being among the most convenient. Although their properties are very well established, linear models are often unrealistic in applications. Moreover, mis-specification of the data generation mechanism by a linear model could lead to large bias. To achieve greater realism, many other parametric models as well as transformation methods have been proposed, each with its own limitations.

Nonparametric modelling makes no assumption on the specification of the model, but it may fail to incorporate some prior information and the resulting estimator of the unknown function tends to incur greater variance. Worse still is the so-called ‘curse of dimensionality’, which renders the standard nonparametric method practically impotent when the dimension of the covariate is high. To ameliorate the ‘curse of dimensionality’, many methods have been proposed to reduce dimension, which includes the projection pursuit (Huber, 1985), the sliced inverse regression (Li, 1991), the single index models (Härdle and Stoker, 1990) and others. There, the model takes the following basic form

$$y=f(X^{\mathrm{T}}\beta_1,\cdots,X^{\mathrm{T}}\beta_q,\varepsilon). \tag{1.1}$$

*The work is supported by the NSF grants DMS-0532370 and NIH grant R01-GM072611.

where y is a response variable, X a p dimensional covariate, ε a random error, and q an integer, which, it is hoped, is much smaller than p. However, model (1.1) has its own limitations. When q is large, the 'curse of dimensionality' remains. Actually, (1.1) is not very practical if the sample size is moderate and q is larger than 2. The interpretability of the model can also arise.

An alternative approach is to relax the conditions imposed on traditional parametric models and explore the hidden structure. Examples include additive models (Breiman and Friedman, 1985; Hastie and Tibshirani, 1990), varying coefficient models (Hastie and Tibshirani, 1993; Fan and Zhang, 1999, 2000; Chiang *et al.*2001), low-dimensional interaction models (Friedman 1991, Gu and Wahba, 1992, Stone *et al.*1997), partially linear models (Wahba 1984; Green and Silverman 1994), and their hybrids (Carroll *et al.*1997, Fan *et al.*1998, Heckman *et al.* 1998, Fan *et al.*2003), among others.

Among the above semiparametric models, the varying coefficient models arise in many contexts. They have been successfully applied to multi-dimensional nonparametric regression, generalized linear models, nonlinear time series models, analysis of longitudinal, functional, and survival data, and financial and economic data.

1.2 Practical meaning

The varying coefficient models are not stimulated by the desire of purely mathematical extension, rather they come from the need in practice. In many scientific areas where statistics is needed, there are some commonly used traditional parametric models found by the people in the area in the light of their experience. Those models are rational in some sense. However, most of them ignore the dynamic feature which may exist in the data set, although the exploration of such dynamic feature sometimes can be very compelling. To explore the dynamic feature and make the model fit the data better, we need to reconsider the modelling strategy. It would not be wise to completely abandon the existing models. It would probably be more sensible to just let the constant parameters evolve with certain characteristics, which leads to the varying coefficient models. For example, to analyse cross-country growth, linear model assumptions are made in the standard growth analyses. However, these assumptions are not supported by the data since the relationship between a set of controls and a particular country's growth rate will depend on its state of development, and the dynamical pattern of this relationship is of importance. It would make much more sense to treat the parameters of growth equations as functions of the state of development, which leads to a standard varying coefficient model. Another example is the analysis of infant mortality in China. The commonly-used model for the analysis of mortality is logistic regression model. Yet, the impacts of the factors on mortality remain constant over time in the model. It is well known that China has been changing dramatically since 1949. It would be implausible to assume the impacts of the factors are constant. They must vary with time, and the dynamic patterns of these impacts are of importance to social studies. Cheng and Zhang (2007) studied the infant mortality data in China, and found the impacts were indeed varying with time. So, it is more sensible to change the constant coefficients in the logistic regression to functional coefficients, which leads to generalized varying coefficient models. The final example is about the circulatory and respiratory problems in Hong Kong. What is interesting is how some environmental factors affect the circulatory and respiratory problems, how the impacts of these factors vary with time. Fan and Zhang (1999) studied this problem very carefully. Applying varying coefficient models, they found the dynamic patterns of the impacts. We will give more detailed description on this effect later.

1.3 Role in the development of statistical methodology

Varying coefficient models are basically locally parametric models. The computation involved in the estimation is cheap and simple: Any existing software for parametric models can easily

be adapted to the need of fitting varying coefficient models. They can be used as trial models to test the efficiency or validity of new statistical methodology developed. For example, for parametric setting, it is well known that in hypothesis test the asymptotic distribution of the maximum likelihood ratio test statistic under null hypothesis does not depend on the nuisance parameters involved in the null hypothesis. This is the so called Wilks phenomenon. Naturally, people would ask whether the Wilks phenomenon still holds for nonparametric setting. Fan *et al.*(2001) have systematically studied this question. They found maximum likelihood ratio test statistics in general may not exist in nonparametric setting. Even if they exist, they would not be optimal. They then introduced the generalized likelihood ratio statistics to overcome the drawbacks of nonparametric maximum likelihood ratio test. They proved that the Wilks phenomenon holds for their generalized likelihood ratio statistics in nonparametric setting. This is a very important finding. The importance lies not only on the elegance of its mathematical beauty but also the practical usage. One straightforward application of this finding is to estimate the distributions of the test statistics under null hypothesis. When sample size is moderate bootstrap method usually outperforms the asymptotic distribution based method. However, the nuisance parameters involved in the null hypothesis have to be evaluated when generating bootstrap samples. How to evaluate the nuisance parameters is the first question one would come up against when using bootstrap. Thanks the Wilks phenomenon, people can just simply assign some reasonable values to the nuisance parameters when generating bootstrap sample to estimate the distribution of the generalized likelihood ratio statistic under null hypothesis. The varying coefficient models as trial models play a very important part in the development of the generalized maximum likelihood ratio test, see Fan *et al.*(2001).

From Sections 1.1, 1.2 and 1.3, we can see that on application side the varying coefficient models are very useful tool to explore the dynamic pattern in many scientific areas, such as economics, finance, politics, epidemiology, medical science, ecology and so on. On theoretical side, they are very useful semiparametric models to get around 'curse of dimensionality'. They are also very nice trial models for the development of new statistical methodology. In the past ten years, the varying coefficient models have seen deep and exciting development. In this paper, we are going to review the major developments on the methodological side of the varying coefficient models.

2 Varying coefficient models

The varying coefficient models are introduced by Cleveland, Grosse and Shyu (1991) to extend the applications of local regression techniques from one-dimensional to multi-dimensional setting. Consider multivariate predictor variables, containing a scalar U and a vector $X = (x_1, \cdots, x_p)^{\mathrm{T}}$.

The varying-coefficient models assume the form of multivariate regression function as

$$m(U,X)=X^{\mathrm{T}}\mathbf{a}(U), \tag{2.1}$$

for unknown functional coefficient $\mathbf{a}(U) = (a_1(U), \cdots, a_p(U))^T$, where $m(U, X) = E(y|U, X)$ is the regression function. An extension of the local regression was given by Hastie and Tibshirani (1993).

In addition to the importance, mentioned in Sections 1.1, 1.2, 1.3, of the varying coefficient models, from statistical modelling point of view, another advantage of the varying coefficient models is that they allow the coefficients to vary smoothly over the group stratified by U and hence permits nonlinear interactions between U and X.

From statistical modelling point of view, the variable U in the varying coefficient models (2.1) may not necessarily be a single variable. Fan, Yao and Cai (2003) proposed an adaptive varying-coefficient model in which $U = X^{\mathrm{T}}\boldsymbol{\beta}$, and $\boldsymbol{\beta}$ was selected by a data driven algorithm.

Throughout this paper, we use $f(u)$ to denote the density function of U, $e_{k,m}$ the unit vector of length m with the k-th component being 1. For any function/functional vector $g(u)$, we use $g^{(k)}(u)$ to denote the kth, $k \geq 2$, derivative of $g(u)$ with respect to u, and $\dot{g}(u)$ the first derivative. We also use $\mathbf{0}_{p\times q}$ to denote a $p \times q$ matrix with each entry being 0, and set $\mu_i = \int u^i K(u)du$ and $\nu_i = \int u^i K^2(u)du$.

2.1 Estimation methods

There are three approaches to estimate the $\mathbf{a}(\cdot)$ in model (2.1). One is kernel-local polynomial smoothing, see Wu *et al.*(1998),Hoover *et al.*(1998),Fan and Zhang (1999),Kauermann and Tutz (1999). One is polynomial spline, see Huang *et al.*(2002,2004) and Huang and Shen (2004). The last one is smoothing spline, see Hastie and Tibshirani (1993),Hoover *et al.* (1998) and Chiang *et al.*(2001). The varying coefficient models, as they stand, are locally linear models. It is more reasonable to use the kernel smoothing method to estimate. In the following, we are going to outline the kernel-local polynomial smoothing method.

2.1.1 Estimation of the functional coefficient—

Suppose that we have a sample $(U_i, X_i^{\mathrm{T}}, y_i)$, $i = 1, \cdots, n$, from (U, X^{T}, y).

$$y = X^{\mathrm{T}}\mathbf{a}(U) + \varepsilon,$$

with $E(\varepsilon) = 0$, and $\mathrm{var}(\varepsilon) = \sigma^2(U)$. For each given u, the local linear estimator $\hat{\mathbf{a}}(u)$ of $\mathbf{a}(u)$ is the part corresponding to $\mathbf{a}$ of the minimizer of

$$L(\mathbf{a},\mathbf{b}) = \sum_{i=1}^{n}\left\{y_i - X_i^{\mathrm{T}}\mathbf{a} - X_i^{\mathrm{T}}\mathbf{b}(U_i - u)\right\}^2 K_h(U_i - u), \tag{2.2}$$

where $K_h(t) = K(t/h)/h$, $K(t)$ is a kernel function, usually taken to be the Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$ and h is bandwidth.

Let

$$\mathbf{X} = (X_1, \cdots, X_n)^{\mathrm{T}},\ \mathbf{U}_u = \mathrm{diag}(U_1 - u, \cdots, U_n - u),\ \Gamma_u = (\mathbf{X}, \mathbf{U}_u\mathbf{X}),$$
$$Y = (y_1, \cdots, y_n)^{\mathrm{T}},\ W_u = \mathrm{diag}\,(K_h(U_1 - u), \cdots, K_h(U_n - u))\,.$$

Then, we have

$$\widehat{\mathbf{a}}(u) = (I_p, \mathbf{0}_p)(\Gamma_u^{\mathrm{T}} W_u \Gamma_u)^{-1}\Gamma_u^{\mathrm{T}} W_u Y, \tag{2.3}$$

where I_p is a size p identity matrix, $\mathbf{0}_p$ is a size p matrix with each entry being 0.

The estimator $\hat{\mathbf{a}}(u)$ is a linear estimator of $\mathbf{a}(u)$. It is asymptotically normally distributed.

Theorem 1: *Under the conditions in* Zhang and Lee (2000), *we have*

$$cov^{-1/2}(\widehat{\mathbf{a}}(u))\{\widehat{\mathbf{a}}(u) - \mathbf{a}(u) - bias(\widehat{\mathbf{a}}(u))\} \xrightarrow{D} N(0, I_p),$$

with

$$bias(\widehat{\mathbf{a}}(u))=2^{-1}\mu_2\mathbf{a}^{(2)}(u)h^2,\ cov(\widehat{\mathbf{a}}(u))=\left\{nhf(u)E(XX^T|U=u)\right\}^{-1}\nu_0\sigma^2(u).$$

The conditional bias and variance of the estimators are also derived in Carroll *et al.*(1998) and Fan and Zhang (1999). Furthermore, the distribution of the maximum discrepancy between the estimated coefficients and true coefficients is given by Xia and Li (1999) and Fan and Zhang (2000).

It is very interesting to look into the asymptotic bias and covariance matrix of $\hat{\mathbf{a}}(u)$. If we ignore μ_2 in the asymptotic bias of $\hat{\mathbf{a}}(u)$, the asymptotic bias would be the remainder of the first-order Taylor's expansion of $\mathbf{a}(U)$ at u. This suggests the bias of $\hat{\mathbf{a}}(u)$ purely comes from the approximation error of the linear approximation of $\mathbf{a}(U)$. In the asymptotic covariance matrix of $\hat{\mathbf{a}}(u)$, the $2hf(u)$ is approximately the probability of U falling into the neighbourhood of u with radius h, and $2nhf(u)$ is approximately the expected number of U_i in the neighbourhood of u. If the kernel function is taken to be the uniform kernel $K(t) = 0.5I(|t|< 1)$, ν_0 would be 0.5, and asymptotic covariance matrix of $\hat{\mathbf{a}}(u)$ would be exactly the covariance matrix of the least squares estimator of the linear model fitting the data in the neighbourhood of u only.

2.1.2 Estimation of bias and variance—Bandwidth selection is an important issue in kernel smoothing. The basic idea of a data driven bandwidth selection procedure is to find an estimator of mean squared error (MSE) of $\hat{\mathbf{a}}(u)$ first, then minimize MSE with respect to bandwidth. The optimal bandwidth is the one minimizing the MSE. To get the estimator of the MSE of $\hat{\mathbf{a}}(u)$, we only need to get the estimator of the bias of $\hat{\mathbf{a}}(u)$ and covariance matrix of $\hat{\mathbf{a}}(u)$. So, it is of importance to estimate the bias and covariance matrix of $\hat{\mathbf{a}}(u)$. In addition to the estimation of MSE, the estimation of bias and covariance matrix are also very important in many other aspects such as hypothesis test and confidence band. In the following, we will briefly describe how to estimate the bias and covariance matrix. It follows the pre-asymptotic substitution idea of Fan and Gijbels (1995).

Let $\mathcal{D}=\left(U_1,X_1^{\mathrm{T}},\cdots,U_n,X_n^{\mathrm{T}}\right)$. By Taylor's expansion and simple calculation, we have

$$E(\widehat{\mathbf{a}}(u)|\mathcal{D})-\mathbf{a}(u)\approx(I_p,\mathbf{0}_p)\left(\Gamma_u^{\mathrm{T}}W_u\Gamma_u\right)^{-1}\Gamma_u^{\mathrm{T}}W_u\tau,$$

where the ith element of $\boldsymbol{\tau}$ is

$$2^{-1}X_i^{\mathrm{T}}\left\{\mathbf{a}^{(2)}(u)(U_i-u)^2+3^{-1}\mathbf{a}^{(3)}(u)(U_i-u)^3\right\}.$$

This naturally leads to the estimator of the conditional bias of $\hat{\mathbf{a}}(u)$ given $\mathcal{D}$

$$\widehat{\mathrm{bias}}(\widehat{\mathbf{a}}(u)|\mathcal{D})=(I_p,\mathbf{0}_p)\left(\Gamma_u^{\mathrm{T}}W_u\Gamma_u\right)^{-1}\Gamma_u^{\mathrm{T}}W_u\widehat{\tau},$$

where $\hat{\boldsymbol{\tau}}$ is $\boldsymbol{\tau}$ with $\mathbf{a}^{(k)}(u)$ being replaced by its estimator $\hat{\mathbf{a}}^{(k)}(u)$, $k = 2, 3$. The estimator $\hat{\mathbf{a}}^{(k)}(u)$, $k = 2, 3$, can be obtained by local cubic fitting with an appropriate pilot bandwidth h_*.

From (2.3), it can be seen

$$\mathrm{cov}(\widehat{\mathbf{a}}(u)|\mathcal{D})=(I_p,\mathbf{0}_p)\left(\Gamma_u^{\mathrm{T}}W_u\Gamma_u\right)^{-1}\left(\Gamma_u^{\mathrm{T}}W_u^2\Gamma_u\right)\left(\Gamma_u^{\mathrm{T}}W_u\Gamma_u\right)^{-1}(I_p,\mathbf{0}_p)^{\mathrm{T}}\sigma^2(u),$$

which leads to the estimator of the conditional covariance matrix of $\hat{\mathbf{a}}(u)$ given $\mathcal{D}$

$$\widehat{\mathrm{cov}}(\widehat{\mathbf{a}}(u)|\mathcal{D}) \approx (I_p,\mathbf{0}_p)\left(\Gamma_u^{\mathrm{T}}W_u\Gamma_u\right)^{-1}\left(\Gamma_u^{\mathrm{T}}W_u^2\Gamma_u\right)\left(\Gamma_u^{\mathrm{T}}W_u\Gamma_u\right)^{-1}(I_p,\mathbf{0}_p)^{\mathrm{T}}\widehat{\sigma}^2(u).$$

The estimator $\hat{\sigma}^2(u)$ can be obtained as a byproduct when we use local cubic fitting with a pilot bandwidth h_* to estimate $\mathbf{a}^{(k)}(u)$, $k = 2, 3$. It is

$$\widehat{\sigma}^2(u)=\frac{Y^{\mathrm{T}}\left\{W_u^* - W_u^*\Gamma_u^*\left(\Gamma_u^{*\mathrm{T}}W_u^*\Gamma_u^*\right)^{-1}\Gamma_u^{*\mathrm{T}}W_u^*\right\}Y}{\mathrm{tr}\left\{W_u^* - (\Gamma_u^{*\mathrm{T}}W_u^*\Gamma_u^*)^{-1}(\Gamma_u^{*\mathrm{T}}W_u^{*2}\Gamma_u^*)\right\}},$$

where W_u^* is W_u with h replaced by h_*, and

$$\Gamma_u^*=(\mathbf{X},\ \mathbf{U}_u\mathbf{X},\ \mathbf{U}_u^2\mathbf{X},\ \mathbf{U}_u^3\mathbf{X}).$$

Please refer to Zhang and Lee (2000) for more detail about the estimation of bias and covariance matrix.

2.1.3 Bandwidth selection—For kernel smoothing approach, bandwidth selection is an important issue. Larger bandwidth may gain on variance side, but loses on bias side. Smaller bandwidth may gain on bias side, but loses on variance side. How to choose an optimal bandwidth is of importance. Wu *et al.*(1998), Hoover *et al.*(1998) proposed to use cross-validation to select the bandwidth. Zhang and Lee (2000) systematically investigated both variable bandwidth and constant bandwidth selection.

Based on the form of varying coefficient models, it is reasonable to define the mean squared error of $\hat{\mathbf{a}}(\cdot)$ as

$$\mathrm{MSE}(h)=E\left\{X^{\mathrm{T}}\widehat{\mathbf{a}}(U) - X^{\mathrm{T}}\mathbf{a}(U)\right\}^2,$$

where (U, X^{T}) is a random vector which shares the same distribution with (U_1,X_1^{T}), and is independent of $\mathcal{D}$. It can be viewed as the future values of the covariates in the sense of prediction. By a simple calculation, we have

$$\mathrm{MSE}(h)=E\left[\mathbf{B}^{\mathrm{T}}(U)\Omega(U)\mathbf{B}(U)+\mathrm{tr}\{\Omega(U)V(U)\}\right],$$

where

$$\mathbf{B}(U)=\mathrm{bias}(\widehat{\mathbf{a}}(U)|U,\mathcal{D}),\ \Omega(U)=E(XX^{\mathrm{T}}|U),\ V(U)=\mathrm{cov}(\widehat{\mathbf{a}}(U)|U,\mathcal{D}).$$

Note that

$$\mathrm{bias}(\widehat{\mathbf{a}}(U)|U,\mathcal{D})=\mathrm{bias}(\widehat{\mathbf{a}}(u)|D)|_{u=U},\ \mathrm{cov}(\widehat{\mathbf{a}}(U)|U,\mathcal{D})=\mathrm{cov}(\widehat{\mathbf{a}}(u)|\mathcal{D})|_{u=U}.$$

For each i, we delete the ith observation, and apply the estimation procedures in Sections 2.1.1 and 2.1.2 to estimate $\mathbf{a}(U_i)$ and the bias and covariance matrix of the estimator of $\mathbf{a}(U_i)$ based on the rest observations. Denote the resulting estimators of bias and covariance matrix of the estimator of $\mathbf{a}(U_i)$ by

$$\widehat{\mathbf{B}}(U_i)=\widehat{\mathrm{bias}}(\widehat{\mathbf{a}}^{\setminus i}(U_i)|\mathcal{D}),\ \widehat{V}(U_i)=\widehat{\mathrm{cov}}(\widehat{\mathbf{a}}^{\setminus i}(U_i)|\mathcal{D}).$$

Let

$$\widehat{\Omega}(U_i)=\frac{\sum_{1\le j\le n,\ j\neq i} X_j X_j^{\mathrm{T}} K_{h_*}(U_j-U_i)}{\sum_{1\le j\le n,\ j\neq i} K_{h_*}(U_j-U_i)}.$$

MSE(h) can be estimated by

$$\widehat{\mathrm{MSE}}(h)=n^{-1}\sum_{i=1}^{n}\left[\widehat{\mathbf{B}}^{\mathrm{T}}(U_i)\widehat{\Omega}(U_i)\widehat{\mathbf{B}}(U_i)+\mathrm{tr}\left\{\widehat{\Omega}(U_i)\widehat{V}(U_i)\right\}\right].$$

The pilot bandwidth h_* for estimating bias and covariance matrix can be chosen by the residual squares criterion (RSC) proposed by Fan and Gijbels (1995). The optimal bandwidth is the one minimizing $\widehat{\mathrm{MSE}}(h)$.

For longitudinal data, it is better to delete the whole *i*th subject rather than just *i*th observation when estimating MSE(h).

The selection of smoothing parameter issue remains for the other two approaches. For polynomial spline approach, the number of knots can be chosen by some commonly used criteria such as CV, AIC, AIC$_c$, BIC and MCV, see Huang *et al.*(2002, 2004), and Huang and Shen (2004). Huang and Shen (2004) also studied how to place the knots. The smoothing parameter with the smoothing spline approach can be selected by cross-validation, see Hoover *et al.*(1998), Chiang *et al.*(2001).

2.1.4 Two-steps estimation—There is an interesting issue arising from the estimation. When the components of **a**(·) have different degrees of smoothness, how to estimate **a**(·)? Intuitively, the smoother components need larger bandwidth whilst the less smooth components need smaller bandwidth. This means it is impossible to optimally estimate all components simultaneously with a single choice of the bandwidth. Indeed, Fan and Zhang (1999) have proved the estimation introduced in Section 2.1.1 (one-step estimation) can not optimally estimate the smoother components no matter how to choose the bandwidth. They then proposed a two-steps idea to estimate the smoother components of **a**(·). They have shown that their proposed two-steps estimation always outperforms one step estimation when estimating the smoother components. They have also shown that two-steps estimation and one-step estimation work equally well when estimating the less smooth components. There is no harm to appeal two-steps estimation scheme.

The idea behind Fan and Zhang's two-steps estimation is to use a smaller bandwidth first to get an initial estimator of the functional coefficient **a**(·). This initial estimator would have larger variance but smaller bias. We then replace the less smooth components of **a**(·) by their initial estimators, and apply higher order smoothing with a slightly larger bandwidth to get the final estimator of the smoother components. The core idea here is that the variance can be reduced by further smoothing, but bias can not be reduced by any kind of smoothing. This is why we have to use a smaller bandwidth in the first step to get an initial estimator with smaller bias.

The two-steps estimation can be sketched as follows. Let $X_i = (x_{i1}, \cdots, x_{ip})^{\mathrm{T}}$ and $\mathbf{a}(\cdot) = (a_1(\cdot), \cdots, a_p(\cdot))^{\mathrm{T}}$. Without loss of generality, we assume $a_p(\cdot)$ is smoother than any $a_j(\cdot)$, $j = 1, \cdots, p-1$, which have the same degree of smoothness. To mathematically formulate it, we assume $a_j(\cdot)$, $j = 1, \cdots, p-1$, have second derivative, and $a_p(\cdot)$ has fourth derivative. We are aiming to estimate $a_p(\cdot)$. To make the description more clear, we write the varying coefficient models as

$$y_i=\sum_{j=1}^{p-1}a_j(U_i)x_{ij}+a_p(U_i)x_{ip}+\varepsilon_i,\ i=1,\cdots,n. \tag{2.4}$$

Applying the estimation introduced in Section 2.1.1 with a smaller bandwidth h, for any given u, we have the initial estimator of $\mathbf{a}(u)$

$$\tilde{\mathbf{a}}(u)=(I_p,\mathbf{0}_p)\left(\Gamma_u^{\mathrm{T}}W_u\Gamma_u\right)^{-1}\Gamma_u^{\mathrm{T}}W_uY.$$

For $j = 1, \cdots, p-1$, replacing $a_j(U_i)$ in model (2.4) by $\tilde{a}_j(U_i)$, the jth component of $\tilde{\mathbf{a}}(U_i)$, we have the synthetic model

$$y_i-\sum_{j=1}^{p-1}\tilde{a}_j(U_i)x_{ij}=a_p(U_i)x_{ip}+\varepsilon_i,\ i=1,\cdots,n. \tag{2.5}$$

As $a_p(\cdot)$ has a fourth derivative, by Taylor's expansion, we have

$$a_p(U_i)\approx\sum_{k=0}^{3}(k!)^{-1}a_p^{(k)}(u)(U_i-u)^k$$

when U_i is in a neighbourhood of u with length $2h_1$. This leads to the following local cubic estimation procedure with bandwidth h_1

$$\sum_{i=1}^{n}\left\{y_i-\sum_{j=1}^{p-1}\tilde{a}_j(U_i)x_{ij}-x_{ip}\sum_{k=0}^{3}a_{p,k}(U_i-u)^k\right\}^2K_{h_1}(U_i-u). \tag{2.6}$$

Minimize (2.6) with respect to $(a_{p,0}, a_{p,1}, a_{p,2}, a_{p,3})$ to get the minimizer. The final estimator of $a_p(u)$ is the part corresponding to $a_{p,0}$ of the minimizer of (2.6), which is

$$\widehat{a}_p(u)=e_{1,4}^{\mathrm{T}}(G^{\mathrm{T}}W_1G)^{-1}G^{\mathrm{T}}W_1\tilde{Y},$$

where $\tilde{Y}=(\tilde{y}_1,\cdots,\tilde{y}_n)^{\mathrm{T}}$, and

$$\tilde{y}_i=y_i-\sum_{j=1}^{p-1}\tilde{a}_j(U_i)x_{ij},\ W_1=\mathrm{diag}\left(K_{h_1}(U_1-u),\cdots,K_{h_1}(U_n-u)\right),$$
$$G=\mathrm{diag}(x_{1p},\cdots,x_{np})Q,\ Q=\begin{pmatrix}1 & U_1-u & (U_1-u)^2 & (U_1-u)^3\\ \vdots & \vdots & \vdots & \vdots\\ 1 & U_n-u & (U_n-u)^2 & (U_n-u)^3\end{pmatrix}. \tag{2.7}$$

Based on the two-steps idea, $a_p(\cdot)$ can also be estimated in another way which is slightly easier to implement. The first step is the same as above, however, in the second step, we just simply smooth $\tilde{a}_p(U_i)$ against U_i by local cubic modelling with bandwidth h_1. For any given u, the resulting final estimator of $a_p(u)$ is

$$\check{a}_p(u)=e_{1,4}^{\mathrm{T}}(Q^{\mathrm{T}}W_1Q)^{-1}Q^{\mathrm{T}}W_1\check{Y},$$

where Q is defined in (2.7), $\check{Y}=(\tilde{a}_p(U_1),\cdots,\tilde{a}_p(U_n))^{\mathrm{T}}$.

The two-steps idea can be widely used in the development of statistical estimation for various models, some complex models in particular. For instance, based on the two-steps idea, Fan and Zhang (2000) developed a novel estimation for longitudinal data analysis. Cheng and Zhang (2007) developed an efficient and easily implemented two-steps estimation for generalized multiparameter likelihood models, and successfully applied it to the analysis of the infant mortality data in China.

2.1.5 Data driven choice of the varying variable—So far, we have assumed that the variable U is known and observable. In an effort to remove this assumption, Fan *et al.*(2003) introduced the following adaptive varying-coefficient model

$$E(Y|X)=\sum_{j=1}^{p} g_j(\beta^T X)x_j, \tag{2.8}$$

where $\boldsymbol{\beta} \in \Re^p$ is an unknown direction, $X = (x_1, \ldots, x_p)^T$. Comparing with the varying coefficient model, $U = \boldsymbol{\beta}^T X$ is an unknown index, including all situations $U = x_1, \cdots, U = x_p$ as specific examples. The identifiably conditions are given in Fan *et al.*(2003). Basically, they shown that the model is identifiably unless

$$E(Y|\mathbf{x})=\alpha^T X \beta^T X+\gamma^T X+c,$$

They proposed an iterative scheme to estimate $\boldsymbol{\beta}$ and the functional coefficients $\{g_j(\cdot)\}$. Given $\boldsymbol{\beta}$, the model (2.8) is really a varying-coefficient model and the functional coefficients can be estimated by using the method in Section 2.1.1, resulting in the estimates $\{\hat{g}(\cdot, \boldsymbol{\beta})\}$. Now, substituting this into (2.8) yields a synthetic parametric model:

$$E(Y|X)=\sum_{j=1}^{p} \widehat{g}_j(\beta^T X,\beta))x_j.$$

The least-squares method can then be applied to estimate $\boldsymbol{\beta}$. Fan *et al.*(2003) gave details on how to implement the estimator, how to select bandwidths, and how to select significant variables in (2.8). They also gave the details on how to extend the techniques to the two-index situations.

2.2 Confidence bands and hypothesis test

2.2.1 Confidence bands—Wu *et al.*(1998) and Chiang *et al.*(2001) studied the pointwise confidence interval for the functional coefficients in varying coefficient models. Wu *et al.* (1998) also investigated the Bonferroni-type confidence bands. For nonparametric inference, the pointwise confidence interval doesn't make much sense. This is because for an unknown function $g(\cdot)$, its $1 - \alpha$ pointwise confidence interval $(g_1(\cdot), g_2(\cdot))$ only guarantees that

$$P=\left(\widehat{g}_1(u) \le g(u) \le \widehat{g}_2(u)\right)=1-\alpha, \text{ for any give } u$$

which does not imply

$$P\left(\widehat{g}_1(u) \le g(u) \le \widehat{g}_2(u), \text{ for any } u \in D\right)=1-\alpha,$$

where D is a compact set.

For nonparametric inference, what is really useful is confidence bands. In the construction of the confidence bands for the functional coefficients in varying coefficient models, the most challenge and important job is to derive the distribution of the maximum discrepancy between

the estimated functional coefficient and true functional coefficient. Fan and Zhang (2000) established the following theorem:

Theorem 2: *Under the conditions in* Fan and Zhang (2000), *we have*

$$P\left\{(-2\log h)^{1/2}\left(\sup_{u\in[0,1]}\frac{|\widehat{a}_j(u)-a_j(u)-\widehat{bias}(\widehat{a}_j(u)|\mathcal{D})|}{\left\{\widehat{var}(\widehat{a}_j(u)|\mathcal{D})\right\}^{1/2}}-d_{v,n}\right)<x\right\}\to\exp\{-2\exp(-x)\},$$

for any given j, j = 1, ⋯, *p, where*

$$d_{v,n}=(-2\log h)^{1/2}+\frac{1}{(-2\log h)^{1/2}}\log\left\{\frac{1}{4\nu_0\pi}\int(K'(t))^2dt\right\},\ \nu_0=\int K^2(t)dt.$$

For any *j, j* = 1, ⋯, *p*, based on Theorem 2, the 1 − *α* confidence bands of $a_j(u)$ can be easily constructed as

$$\widehat{a}_j(u)-\widehat{\text{bias}}(\widehat{a}_j(u)|\mathcal{D})\pm\Delta_{1,\alpha}(u),$$

where

$$\Delta_{j,\alpha}(u)=\left(d_{v,n}+\left[\log 2-\log\{-\log(1-\alpha)\}\right](-2\log h)^{-1/2}\right)\{\widehat{\text{var}}(\widehat{a}_j(u)|\mathcal{D})\}^{1/2}.$$

The estimator $\widehat{\text{bias}}(\widehat{a}_j(u)|\mathcal{D})$ of the conditional bias of $\hat{a}_j(u)$ and the estimator $\widehat{\text{var}}(\widehat{a}_j(u)|\mathcal{D})$ of the conditional variance of $\hat{a}_j(u)$ can be obtained through the estimation introduced in Section 2.1.2. Fan and Zhang (2000) have shown this confidence bands works quite well.

Huang *et al.*(2002, 2004) investigated the pointwise confidence intervals and confidence bands based on polynomial spline approach and the Bonferroni adjustment.

2.2.2 Hypothesis test—In the varying coefficient model (2.1), the inference questions arise naturally such as whether the coefficients are really varying and if certain components of covariates *X* are statistically significant. This amounts to testing

$$H_0:a_j(u)=C_j\leftrightarrow H_1:a_j(u)\neq C_j, \tag{2.9}$$

C_j is a constant.

Cai, Fan and Yao (2000) developed a bootstrap based test for the hypothesis (2.9). The generalized likelihood ratio (GLR) test was developed to address this kind of question. See Fan *et al.*(2001) and Section 3.2. Fan and Zhang (2000) took another approach which was based on the asymptotic distribution of the maximum discrepancy between the estimated functional coefficient and true functional coefficient. They established the following Theorem

Theorem 3: *Under the conditions in* Fan and Zhang (2000), *when* $a_j(\cdot)$ *is a constant* C_j *we have*

$$P(T_j<x)\to\exp\{-2\exp(-x)\},$$

where

$$T_j=(-2\log h)^{1/2}\left(\sup_{u\in[0,1]}|\{\widehat{var}(\widehat{a}_j|\mathcal{D})\}^{-1/2}\left(\widehat{a}_j-\widehat{C}_j-\widehat{bias}(\widehat{a}_j|\mathcal{D})\right)|-d_{v,n}\right),$$

with

$$\widehat{C}_j=n^{-1}\sum_{i=1}^{n}\widehat{a}_j(U_i).$$

Based on Theorem 3, the test for the hypothesis (2.9) can be constructed as rejecting H_0 if the values of the test statistic T_j exceeds the asymptotic critical value $c_\alpha=-\log\{-0.5\log(1-\alpha)\}$.

Taking polynomial spline approach, Huang *et al.*(2002) proposed a goodness-of-fit test for the hypothesis (2.9) based on the comparison of the weighted residual sum of squares. They used the bootstrap to implement their test. It is a specific incidence of GLR studied by Fan *et al.* (2001).

2.3 Semivarying coefficient models

In practice, sometimes, some of the components of $\mathbf{a}(\cdot)$ in model (2.1) are constant, while other components have interactions with U. Without loss of generality, we can write the model as

$$y=Z_1^{\mathrm{T}}\mathbf{a}_1(U)+Z_2^{\mathrm{T}}\mathbf{a}_2+\varepsilon, \tag{2.10}$$

where $(Z_1^{\mathrm{T}},Z_2^{\mathrm{T}})^{\mathrm{T}}=X$. Z_i is a p_i dimensional covariate, $i=1, 2$, and $p_1+p_2=p$. Model (2.10) cannot be treated statistically as a special case of varying coefficient models, as the information that $\mathbf{a}_2$ is a constant vector should be fully utilized. Zhang *et al.*(2002) studied the semivarying coefficient model (2.10). They proposed a two-steps estimation procedure, and showed their estimator of $\mathbf{a}_2$ is of convergence rate $O_P(n^{-1/2})$, and their estimator of $\mathbf{a}_1(\cdot)$ is as well as when $\mathbf{a}_2$ is known. For model (2.10), the estimation of $\mathbf{a}_2$ is of most interest. Because, a good estimator $\hat{\mathbf{a}}_2$ of $\mathbf{a}_2$ should be of convergence rate $O_P(n^{-1/2})$. After substituting $\hat{\mathbf{a}}_2$ for the $\mathbf{a}_2$ in model (2.10), (2.10) would become a standard varying coefficient models. As $\hat{\mathbf{a}}_2$ is of convergence rate $O_P(n^{-1/2})$, the substitution $\hat{\mathbf{a}}_2$ for $\mathbf{a}_2$ would have little influence on the estimation of the functional coefficient $\mathbf{a}_1(\cdot)$. So, the standard estimation for standard varying coefficient models such as the one in Section 2.1.1 can be applied to estimate $\mathbf{a}_1(\cdot)$.

The estimation of $\mathbf{a}_2$ in Zhang *et al.*(2002) can be briefed as follows: We first treat $\mathbf{a}_2$ as functional, and appeal the estimation in Section 2.1.1 to get an initial estimator of $\mathbf{a}_2(U_i)$, $i=1, \cdots, n$,

$$\tilde{\mathbf{a}}_2(U_i)=(\mathbf{0}_{p_2\times p_1},I_{p_2},\mathbf{0}_{p_2\times p})\left(\Gamma_{U_i}^{\mathrm{T}}W_{U_i}\Gamma_{U_i}\right)^{-1}\Gamma_{U_i}^{\mathrm{T}}W_{U_i}Y.$$

Then, we average $\tilde{\mathbf{a}}_2(U_i)$ over $i=1, \cdots, n$ to get the final estimator of $\mathbf{a}_2$

$$\widehat{\mathbf{a}}_2=n^{-1}\sum_{i=1}^{n}(\mathbf{0}_{p_2\times p_1},I_{p_2},\mathbf{0}_{p_2\times p})\left(\Gamma_{U_i}^{\mathrm{T}}W_{U_i}\Gamma_{U_i}\right)^{-1}\Gamma_{U_i}^{\mathrm{T}}W_{U_i}Y.$$

Intuitively, the covariance matrix of $\hat{\mathbf{a}}_2$ should be of order $O(n^{-1})$ because for each i, $\tilde{\mathbf{a}}_2(U_i)$ is obtained locally around U_i and $(U_i,X_i^{\mathrm{T}},y_i)$, $i=1, \cdots, n$, are independent with each other. Indeed, Zhang *et al.*(2002) showed the conditional bias of the estimator $\hat{\mathbf{a}}_2$ given $\mathcal{D}$ is of order $O_P(h^2)$, and the conditional covariance matrix of $\hat{\mathbf{a}}_2$ given $\mathcal{D}$ is of order $O_P(n^{-1})$ under some regularity conditions. This implies that when the bandwidth for the initial estimator $\tilde{\mathbf{a}}_2(U_i)$ in

the first step is taken to be of order $O(n^{-1/4})$, the estimator $\hat{\mathbf{a}}_2$ would have convergence rate $O_P(n^{-1/2})$.

Although Zhang *et al.*'s estimation of $\mathbf{a}_2$ is easy to implement and the resulting estimator has convergence rate of order $O_P(n^{-1/2})$, the asymptotic varaince of the estimator does not reach the lower bound for the semiparametric model.

Fan and Huang (2005) have more deeply investigated model (2.10). They proposed a profile least-squares technique to estimate $\mathbf{a}_2$, and established the asymptotic normality of the estimator. They also introduced the profile likelihood ratio test and demonstrated that the test statistic followed asymptotically χ^2 distribution under null hypothesis which unveiled a new Wilks type of phenomenon.

Fan and Huang's profile least-squares estimation can be outlined as follows: We first pretend $\mathbf{a}_2$ is known, and write the model (2.10) as

$$y_i - Z_{i2}^{\mathrm{T}}\mathbf{a}_2 = Z_{i1}^{\mathrm{T}}\mathbf{a}_1(U_i) + \varepsilon_i,\ i = 1, \cdots, n, \tag{2.11}$$

where $(Z_{i1}^{\mathrm{T}}, Z_{i2}^{\mathrm{T}})^{\mathrm{T}} = X_i$. Applying the estimation in Section 2.1.1, we get the estimator of $\mathbf{a}_1(U_i)$

$$\tilde{\mathbf{a}}_1(U_i) = (I_{p_1}, \mathbf{0}_{p_1})\left(\tilde{\Gamma}_{U_i}^{\mathrm{T}} W_{U_i} \tilde{\Gamma}_{U_i}\right)^{-1} \tilde{\Gamma}_{U_i}^{\mathrm{T}} W_{U_i}\ \tilde{Y},$$

where $\tilde{\Gamma}_u$ is the Γ_u with $\mathbf{X}$ replaced by $\mathbf{Z}_1$,

$$\mathbf{Z}_1 = (Z_{11}, \cdots, Z_{n1})^{\mathrm{T}},\quad \tilde{Y} = Y - \mathbf{Z}_2\mathbf{a}_2,\ \mathbf{Z}_2 = (Z_{12}, \cdots, Z_{n2})^{\mathrm{T}}.$$

Substituting $\tilde{\mathbf{a}}_1(U_i)$ for $\mathbf{a}_1(U_i)$ in model (2.11), we have the following synthetic model

$$y_i - (Z_{i1}^{\mathrm{T}}, \mathbf{0}_{1\times p_1})\left(\tilde{\Gamma}_{U_i}^{\mathrm{T}} W_{U_i} \tilde{\Gamma}_{U_i}\right)^{-1} \tilde{\Gamma}_{U_i}^{\mathrm{T}} W_{U_i}\ \tilde{Y} = Z_{i2}^{\mathrm{T}}\mathbf{a}_2 + \varepsilon_i,\ i = 1, \cdots, n,$$

which can be written in maxtrix form as

$$(I_n - \mathbf{S})Y = (I_n - \mathbf{S})\mathbf{Z}_2\mathbf{a}_2 + \varepsilon, \tag{2.12}$$

where

$$\mathbf{S} = \begin{pmatrix} (Z_{11}^{\mathrm{T}}, \mathbf{0}_{1\times p_1})\left(\tilde{\Gamma}_{U_1}^{\mathrm{T}} W_{U_1} \tilde{\Gamma}_{U_1}\right)^{-1} \tilde{\Gamma}_{U_1}^{\mathrm{T}} W_{U_1} \\ \vdots \\ (Z_{n1}^{\mathrm{T}}, \mathbf{0}_{1\times p_1})\left(\tilde{\Gamma}_{U_n}^{\mathrm{T}} W_{U_n} \tilde{\Gamma}_{U_n}\right)^{-1} \tilde{\Gamma}_{U_n}^{\mathrm{T}} W_{U_n} \end{pmatrix},\ \text{and}\ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Appealing least squares estimation to model (2.12), we get the estimator of $\mathbf{a}_2$

$$\widehat{\mathbf{a}}_2 = \left\{\mathbf{Z}_2^{\mathrm{T}}(I_n - \mathbf{S})^{\mathrm{T}}(I_n - \mathbf{S})\mathbf{Z}_2\right\}^{-1}\mathbf{Z}_2^{\mathrm{T}}(I_n - \mathbf{S})^{\mathrm{T}}(I_n - \mathbf{S})Y. \tag{2.13}$$

Fan and Huang (2005) have shown the covariance matrix of $\hat{\mathbf{a}}_2$ given in (2.13) reaches the lower bound for semiparametric models.

Theorem 4—*Under the conditions in* Fan and Huang (2005), *the estimator* $\hat{\mathbf{a}}_2$ *given in (2.13) is asymptotic normal, i.e.*

$$n^{1/2}(\widehat{\mathbf{a}}_2 - \mathbf{a}_2) \xrightarrow{D} N(0, \sum),$$

where

$$\sum = \left(E(\mathbf{Z}_2\mathbf{Z}_2^T) - E\left[E(\mathbf{Z}_2\mathbf{Z}_1^T|U)\left\{E(\mathbf{Z}_1\mathbf{Z}_1^T|U)\right\}^{-1} E(\mathbf{Z}_1\mathbf{Z}_2^T|U)\right]\right)^{-1} \sigma^2(u).$$

It can be shown $\boldsymbol{\Sigma}$ is a semiparametric efficient bound for semivarying coefficient models when $\varepsilon \sim N(0, \sigma^2)$. Hence the profile least squares estimator of $\mathbf{a}_2$ is semiparametrically efficient.

Ahmad *et al.*(2005) used a general series method to estimate the semivarying coefficient model (2.10). Xia, Zhang and Tong (2004) proposed a cross-validation based model selection procedure to find which components are constant and which are functional in practice. This can also be done by the GLR test of Fan *et al.*(2001). Li and Liang (2007) studied the variable selection issue with the semivarying coefficient models, and Fan and Huang (2005) studied the inference of parametric part $\mathbf{a}_2$ in the semi-varying modeling using GLR test.

2.4 The circulatory and respiratory problems in Hong Kong

We now briefly illustrate the standard varying coefficient models via an application to an environmental data set. The data set used here consists of a collection of daily measurements of pollutants and other environmental factors in Hong Kong between January 1, 1994 and December 31, 1995 (Courtesy of Professor T.S. Lau). Three pollutants, Sulphur Dioxide (in $\mu g/m^3$), Nitrogen Dioxide (in $\mu g/m^3$) and Dust (in $\mu g/m^3$), are considered here.

An objective of the study is to understand the association between level of the pollutants and number of daily total hospital admissions for circulatory and respiratory problems and to examine the extent to which the association varies over time.

We consider relationship among the number of daily hospital admission (y) and level of pollutants Sulphur Dioxide, Nitrogen Dioxide and Dust, which are denoted by x_2, x_3 and x_4, respectively. We took $x_1 = 1$ − the intercept term, and U = time. The varying-coefficient model

$$y = a_1(U) + a_2(U)x_2 + a_3(U)x_3 + a_4(U)x_4 + \varepsilon \tag{2.14}$$

is used to fit the data set.

As the two-steps estimation procedure stated in section 2.1.4 is better than one-step estimation, the two-steps estimation procedure is employed to estimate the functional coefficients in (2.14). Indeed, from the estimated functional coefficients, see Fig. 1, we can see the functional coefficients have different degrees of smoothness, the two-steps estimation is necessary for this data set.

Fig. 1 depicts the estimated functional coefficients. They describe the extent to which the coefficients vary with time. Two short dashed curves indicate 95% confidence intervals with bias ignored. The standard errors are computed based on the local cubic regression in the second step. See Section 4.3 of Fan and Gijbels (1996) on how to compute the estimated standard errors for the univariate local polynomial regression. The figure shows that there is strong time effect on the coefficients, which suggests the impacts of the three pollutants concerned on the circulatory and respiratory problems do vary with time.

Given the type of this paper, we have not gone the details of the analysis for this data set. For rigorous analysis for this data set, please refer to Fan and Zhang (1999).

3 Generalized varying-coefficient models

The varying-coefficient models can readily be extended to the exponential-family of conditional distributions. This allows us to more effectively deal with various types of response variables. Via the canonical link function $g(\cdot)$, the regression function is modeled as

$$g(m(U,X))=\theta(U,X)=X^{\mathrm{T}}\mathbf{a}(U). \tag{3.1}$$

Here, X is still a p dimensional covariate, and U a covariate of scalar.

To make our methodology and theory more general, we do not confine our discussion in the exponential-family. Rather we only assume the log conditional density function (we define density function as the probability function when y is discrete) of y given (U, X^{T}) is $\ell(m(U, X), y)$.

3.1 Estimation procedure

Among various estimation methods, the local maximum likelihood estimation seems more natural and reasonable to estimate the functional coefficient $\mathbf{a}(\cdot)$ in the generalized varying coefficient models. The local maximum likelihood estimation can be briefly described as follows.

Denote $\dot{\mathbf{a}}(u)$ by $\mathbf{b}(u)$. For any given u, the local maximum likelihood estimator $(\hat{\mathbf{a}}^{\mathrm{T}}(u), \hat{\mathbf{b}}^{\mathrm{T}}(u))$ of $(\mathbf{a}^{\mathrm{T}}(u), \mathbf{b}^{\mathrm{T}}(u))$ is the maximizer of the local log-likelihood function

$$\mathcal{L}(\mathbf{a},\mathbf{b})=\sum_{i=1}^{n}\ell\left(g^{-1}\left[X_i^{\mathrm{T}}\{\mathbf{a}+\mathbf{b}(U_i-u)\}\right],y_i\right)K_h(U_i-u). \tag{3.2}$$

Cai *et al.*(2000) have established the asymptotic normality of the local maximum likelihood estimator of $\mathbf{a}(u)$.

Theorem 5—*Under the conditions in* Cai et al.(2000)*, we have*

$$(nhf(u)/\nu_0)^{1/2}\left\{\widehat{\mathbf{a}}(u)-\mathbf{a}(u)-2^{-1}\mu_2\mathbf{a}^{(2)}(u)h^2\right\}\xrightarrow{D}N(\mathbf{0}_{p\times 1},\ \sum),$$

where

$$\sum=\left(E\left[E\left\{\frac{\partial^2\ell(g^{-1}(X^T\mathbf{a}(u)),y)}{\partial(X^T\mathbf{a}(u))^2}|X,U\right\}XX^T|U=u\right]\right)^{-1}.$$

From Theorem 5, we can see, the bias of $\hat{\mathbf{a}}(u)$ is the same as that in standard varying coefficient models. As we see before, the $nhf(u)/\nu_0$ in the asymptotic covariance matrix is the expected number of data in the neighbourhood of u with the length $2h$, and the $\mathbf{\Sigma}$ in the asymptotic covariance matrix is like the Fisher information matrix in parametric setting. Theorem 5 is like the local version of the asymptotic normality of maximum likelihood estimator in parametric setting. Base on Theorem 5, it is obvious that the local maximum likelihood estimation is efficient.

If the conditional distribution of y given U and X belongs to the exponential-family, and the link function $g(\cdot)$ is canonical link, the $\mathbf{\Sigma}$ in Theorem 5 can be simplified to $E\{XX^{\mathrm{T}}\mathrm{var}(y|U, X)|U = u\}$.

Although the local maximum likelihood estimation is efficient, it could be difficult to implement as the local maximum likelihood estimator in general does not have a closed form. It is often computationally expensive to maximize (3.2). Instead of maximizing (3.2), Cai *et al.*(2000) proposed a one-step Newton-Raphson estimation of $\mathbf{a}(u)$ to ease the computational burden.

Let $\mathcal{L}(\mathbf{a}, \mathbf{b})$ and $\ddot{\mathcal{L}}(\mathbf{a}, \mathbf{b})$ be the first and second derivative of $\mathcal{L}(\mathbf{a}, \mathbf{b})$ respectively. Denote $(\mathbf{a}^{\mathrm{T}}(u), \mathbf{b}^{\mathrm{T}}(u))^{\mathrm{T}}$ by $\boldsymbol{\beta}(u)$. Let $(\mathbf{a}_0^{\mathrm{T}},\mathbf{b}_0^{\mathrm{T}})^{\mathrm{T}}$ be the initial estimator of $\boldsymbol{\beta}(u)$. The one-step Newton-Raphson estimator of $\boldsymbol{\beta}(u)$ is

$$\widehat{\beta}_{os}(u)=(\mathbf{a}_0^{\mathrm{T}},\mathbf{b}_0^{\mathrm{T}})^{\mathrm{T}} - \ddot{\mathcal{L}}(\mathbf{a}_0,\mathbf{b}_0)^{-1}\ \dot{\mathcal{L}}\ (\mathbf{a}_0,\mathbf{b}_0). \tag{3.3}$$

The one-step Newton-Raphson estimation can be implemented in the following way: Suppose we wish to evaluate the function $\hat{\mathbf{a}}(\cdot)$ at grid points u_j, $j = 1, \cdots, m$. Pinch the central point u_{i_0}, $i_0 = m/2$, compute the local maximum likelihood estimator $\hat{\boldsymbol{\beta}}(u_{i_0})$, use this estimator as initial estimator at u_{i_0+1}, and apply (3.3) to get the estimator $\hat{\boldsymbol{\beta}}_{os}(u_{i_0+1})$. Now use $\hat{\boldsymbol{\beta}}_{os}(u_{i_0+1})$ as initial estimator at point u_{i_0+2}, and apply (3.3) to get $\hat{\boldsymbol{\beta}}_{os}(u_{i_0+2})$ and so on. Likewise, we can compute $\hat{\boldsymbol{\beta}}_{os}(u_{i_0-1})$, $\hat{\boldsymbol{\beta}}_{os}(u_{i_0-2})$ and so on. In this way, we obtain our estimates at all grid points.

Cai *et al.*(2000) showed that the one-step Newton-Raphson estimator can save computational cost in an order of tens without deteriorating its performance.

An interesting and important issue with generalized varying coefficient models is how to estimate the covariance matrix of the local maximum likelihood estimator. Cai *et al.*(2000) proposed a sandwich method to estimate the covariance matrix of $\hat{\mathbf{a}}(u)$. According to the sandwich method, the covariance matrix of $\hat{\mathbf{a}}(u)$ can be estimated by,

$$\widehat{\mathrm{cov}}(\widehat{\mathbf{a}}(u))=(I_p,\mathbf{0}_p)\widehat{\Lambda}_2^{-1}\widehat{\Lambda}_1\widehat{\Lambda}_2^{-1}(I_p,\mathbf{0}_p)^{\mathrm{T}}, \tag{3.4}$$

with

$$\widehat{\Lambda}_2=\sum_{i=1}^{n} q_2\left[X_i^{\mathrm{T}}\left(\widehat{\mathbf{a}}(u)+\widehat{\mathbf{b}}(u)(U_i-u)\right),\ y_i\right]\mathbf{H}_i\otimes(X_iX_i^{\mathrm{T}})K_h(U_i-u),$$
$$\widehat{\Lambda}_1=\sum_{i=1}^{n} q_1^2\left[X_i^{\mathrm{T}}\left(\widehat{\mathbf{a}}(u)+\widehat{\mathbf{b}}(u)(U_i-u)\right),\ y_i\right]\mathbf{H}_i\otimes(X_iX_i^{\mathrm{T}})K_h^2(U_i-u),$$

where

$$\mathbf{H}_i=(1,\ U_i-u)^{\mathrm{T}}(1,\ U_i-u),\ q_k(t,y)=(\partial^k/\partial t^k)\ell\left\{g^{-1}(t),\ y\right\}.$$

Cai *et al.*(2000) showed the sandwich method worked quite well by extensive simulation studies. Fan and Peng (2004) have proved the consistency of the sandwich estimator.

The bias and variance of the local maximum likelihood estimator can be estimated by using the general method outlined in Fan *et al.*(1998). In general, it is difficult to accurately estimate the bias of $\hat{\mathbf{a}}(u)$ due to poor estimation of higher order derivative of $\mathbf{a}(u)$. In the construction of confidence bands, an alternative approach to deal with the bias is to use a slightly smaller bandwidth to make the bias ignorable.

Like the standard varying coefficient models, the bandwidth plays a very important role in the local maximum likelihood estimation for generalized varying coefficient models. A natural approach to select the bandwidth is to appeal the cross-validation idea. For each i, we delete the ith observation, and estimate $\mathbf{a}(U_i)$ based on the rest of the observations. Let $\hat{\mathbf{a}}^{\backslash i}(U_i)$ be the obtained estimator. The sum of cross-validation is defined as

$$\mathrm{CV} = -\sum_{i=1}^{n} \ell\left\{g^{-1}\left(X_i^{\mathrm{T}}\hat{\mathbf{a}}^{\backslash i}(U_i)\right),\ y_i\right\}.$$

We compute the CVs for different bandwidths in a reasonable range. The selected bandwidth is the one minimizing the CV.

3.2 Hypothesis test

Like standard varying coefficient models, whether some certain coefficients are really varying with U or whether some certain coefficients are significantly different from 0 is of interest and importance. These questions can be formulated to the hypotheses

$$H_0: a_k(\cdot) = a_k,\ k = 1, \cdots, p, \tag{3.5}$$

and

$$H_0: a_k(\cdot) = 0,\ \text{for certain } k.$$

While these two problems look alike, there are very different statistically. The former tests the parametric null hypothesis against the nonparametric alternative hypothesis, while the latter tests against the nonparametric null hypothesis as the null hypothesis contains unknown nonparametric components $a_j(\cdot)$ for $j \neq k$. Cai *et al.*(2000) discussed how to construct the hypothesis test for these hypotheses based on the generalized maximum likelihood ratio test developed by Fan *et al.*(2001). The generalized maximum likelihood ratio test is easy to implement and has good power. Taking the hypothesis (3.5) as an example, the generalized maximum likelihood ratio test statistic is the difference between the log likelihood functions under the alternative and null hypotheses, which is

$$T = \sum_{i=1}^{n}\left(\ell\left[g^{-1}\left\{X_i^{\mathrm{T}}\hat{\mathbf{a}}(U_i)\right\},\ y_i\right] - \ell\left\{g^{-1}\left(X_i^{\mathrm{T}}\hat{\mathbf{a}}\right),\ y_i\right\}\right),$$

where $\hat{\mathbf{a}}(\cdot)$ is the local maximum likelihood estimator of functional coefficient $\mathbf{a}(\cdot)$ under alternative hypothesis, and $\hat{\mathbf{a}}$ is the maximum likelihood estimator of the constant vector $\mathbf{a} = (a_1, \cdots, a_p)^{\mathrm{T}}$ under null hypothesis. The test is that we reject the null hypothesis when $T > c_\alpha$, where c_α, is the critical value which can be computed by either asymptotic distribution of T or bootstrap under null hypothesis.

The asymptotic distribution of T under null hypothesis is normal distribution and free of the value of $\mathbf{a}$. This is the so called Wilks phenomenon. For more rigorous justification, we refer to the article by Fan *et al.*(2001). When sample size is moderate, it is better to use bootstrap under null hypothesis to estimate the critical value c_α. We have to evaluate $\mathbf{a}$ when generating bootstrap samples under null hypothesis. Thanks the Wilks phenomenon, the distribution of T under null hypothesis is free of the value of $\mathbf{a}$, so we can just simply assign a reasonable value to $\mathbf{a}$. We recommend to evaluate $\mathbf{a}$ by the maximum likelihood estimator $\hat{\mathbf{a}}$ of $\mathbf{a}$ under the null hypothesis.

Finally, we would like to point out that the local maximum likelihood method can be regarded as a specific case of the local estimation equation method of Carroll *et al.*(1998). Kauermann and Tutz (1999) proposed a graphical technique for checking the discrepancy between a parametric model and a varying coefficient model. Qu and Li (2006) investigated a nonparametric goodness of fit test for the generalized varying coefficient models.

4 Analysis of longitudinal and functional data

In many applications, data for different individuals are collected over a period of time. The number of data points for different individuals can be different and so is the location of time. Such a kind of data are called longitudinal data. Often, interest lies in studying the association between the covariates and the response variable. To this end, a linear model is often employed:

$$Y(t)=\beta_0+X(t)^{\mathrm{T}}\beta+\varepsilon(t), \tag{4.1}$$

for covariates and response variable collected at time t. See for example Diggle *et al.*(2002) and Hand and Crowder (1996).

Despite of its success in many applications, model (4.1) does not allow the association to vary over time, even though the covariates and the response variable change over time and environment. To account for this, Zeger and Diggle (1994) proposed a semiparametric model, by allowing the intercept term β_0 to depend on the time, but not the other coefficients. To genuinely examine whether the association changes over time, Brumback and Rice (1998) and Hoover *et al.*(1998) propose the following varying coefficient model

$$Y(t)=\beta_0(t)+X(t)^{\mathrm{T}}\beta(t)+\varepsilon(t), \tag{4.2}$$

where the functional coefficients are assumed to be smooth. The functional coefficients can also be a function of other covariates instead of the time variable. This is a specific case of the functional linear model discussed in Ramsay and Silverman (1997) for functional data analysis. When covariates are absent, model (4.2) was studied by Rice and Silverman (1991) and Hart and Wehrly (1993) for functional data.

The coefficients in model (4.2) can be estimated by the kernel, polynomial and smoothing spline methods (Brumback and Rice, 1998;Hoover *et al.*1998,Huang *et al.*2002,2004). Fan and Zhang (2000) proposed a two-steps method to overcome the computational burden of the smoothing spline methods. The approaches for constructing confidence regions based on the kernel method can be found in Wu and Chiang (2000) and Chiang *et al.*(2001). The construction of confidence bands based on polynomial spline method can be found in Huang *et al.*(2002, 2004).

One important issue with longitudinal data analysis is how to incorporate the within subject correlation structure into the estimation procedure. For parametric setting, this issue has been thoroughly investigated, and the methodology has been well established; see e.g. Diggle *et al.* (2002) and the references therein. The situation with nonparametric based longitudinal data analysis is quite different, see Lin and Carroll (2001).

Various studies have been made on the partial linear model in which the coefficients $\boldsymbol{\beta}(t)$ in (4.2) are constant. Lin and Ying (2001) employed a counting process approach which is ameliorated by Fan and Li (2004). An important discovery made by Lin and Carroll (2001) is that the commonly-used forms of the kernel method are local and can not incorporate the within subject correlation. An innovative kernel method is proposed by Wang (2003), which incorporates the true covariance structure. The idea has been successfully extended to the partial linear model by Wang *et al.*(2003), which achieves the semiparametric efficient bound

computed in Lin and Carroll (2001). Qu and Li (2006) proposed an estimation procedure for the varying coefficient models based on the penalized spline and quadratic inference function approaches. The advantage of Qu and Li's estimation is it can directly incorporate within subject correlation into the estimation without any need to estimate the nuisance parameters associated with the correlation. Despite the need of within subject covariance for longitudinal studies, few studies have been made.

Missing data such as dropout are common in long term longitudinal studies. Hogan *et al.* (2004) studied the mixtures of varying coefficient models for longitudinal data with discrete or continuous nonignorable dropout.

The within subject correlation structure plays a very important role in longitudinal data analysis. This is because not only an estimator can be improved by incorporating the within subject correlation structure into the estimation procedure, but also the within subject correlation structure can sometimes shed valuable insights in practical problems, see Sun *et al.*(2008). Fan *et al.*(2007) and Sun *et al.*(2008) have systematically studied the estimation of the within subject correlation structure. In the following, we shall briefly introduce the estimation of the within subject covariance matrix.

When p_2 in the semivarying coefficient models (2.10) is 0, the semivarying coefficient models becomes varying coefficient models. It is reasonable to view the semivarying coefficient models as an extension of varying coefficient models. We therefore consider the semivarying coefficient models

$$y(t)=Z_1^{\mathrm{T}}(t)\mathbf{a}_1(t)+Z_2^{\mathrm{T}}(t)\mathbf{a}_2+\varepsilon(t), \tag{4.3}$$

where $Z_i(t)$, $i = 1, 2$, is p_i dimensional covariate, $p_1 + p_2 = p$, and $E\varepsilon(t) = 0$, and $\mathrm{var}(\varepsilon(t)) = \sigma^2(t)$ which is unknown. Modeling the covariance matrix is intrinsically challenging due to sparse irregular observed time points for each individual. To take this structure into account, Fan *et al.*(2007) proposed the following semiparametric model. The variance function $\sigma^2(\cdot)$ are modelled nonparametrically and the correlation function parametrically: For any t and s, the correlation between $\varepsilon(t)$ and $\varepsilon(s)$ is $\rho(s, t, \boldsymbol{\theta})$. The function form of $\rho(s, t, \boldsymbol{\theta})$ is known, but $\boldsymbol{\theta}$ is unknown to be estimated. In this way, the variance function $\sigma^2(t)$ is estimable nonparametrically as long as the collection of time points for all individuals are dense in a time interval of interest. On the other hand, sparse individual observations (for those individuals with at least two observations) can be aggregated to estimate the parameters in the correlation function. The idea is indeed powerful and takes longitudinal data structure at heart. The modeling biases of correlation functions can be reduced by expanding the family of parametric functions, such as the linear combinations of the ARMA-correlation and random-effect-correlation structure.

Suppose that a sample from (4.3) consists of n subjects. For each i, $i = 1, \cdots, n$, for the ith subject, we have observation $(Z_{i1}^{\mathrm{T}}(t_{ij}), Z_{i2}^{\mathrm{T}}(t_{ij}), y_i(t_{ij}))$ at time point t_{ij}, $j = 1, \cdots, J_i$. Let $\varepsilon_i(t_{ij})$ be the $\varepsilon(t)$ corresponding to $(Z_{i1}^{\mathrm{T}}(t_{ij}), Z_{i2}^{\mathrm{T}}(t_{ij}), y_i(t_{ij}))$, and $\boldsymbol{\varepsilon}_i = (\varepsilon_i(t_{i1}), \cdots, \varepsilon_i(t_{iJ_i}))^{\mathrm{T}}$. Denote the covariance matrix of $\boldsymbol{\varepsilon}_i$ by Σ_i.

Of interest is estimating model parameters $\mathbf{a}_1(\cdot)$, $\mathbf{a}_2$, $\sigma^2(\cdot)$ and $\boldsymbol{\theta}$. On one hand, the estimation of σ^2 and $\boldsymbol{\theta}$ depends on the estimation of $\mathbf{a}_1(\cdot)$ and $\mathbf{a}_2$. On the other hand, the estimation of $\mathbf{a}_1(\cdot)$ and $\mathbf{a}_2$ can be improved by using the estimate of σ^2 and $\boldsymbol{\theta}$. Therefore, the estimation must be done in steps. The initial estimators of $\mathbf{a}_1(\cdot)$ and $\mathbf{a}_2$ are constructed by ignoring the within subject correlation. With these estimators, we can estimate σ^2 and $\boldsymbol{\theta}$. Finally, we can estimate $\mathbf{a}_1(\cdot)$ and $\mathbf{a}_2$ more efficiently by using the estimators of σ^2 and $\boldsymbol{\theta}$.

Applying the profile least squares estimation in Section 2.3 with weighted least squares estimation for the synthetic linear model (2.12), we have the estimator of $\mathbf{a}_2$

$$\widehat{\mathbf{a}}_2=\left\{\mathbf{Z}_2^{\mathrm{T}}(I_n-\mathbf{S})^{\mathrm{T}}W(I_n-\mathbf{S})\mathbf{Z}_2\right\}^{-1}\mathbf{Z}_2^{\mathrm{T}}(I_n-\mathbf{S})^{\mathrm{T}}W(I_n-\mathbf{S})Y,$$

where $\mathbf{S}$, $\mathbf{Z}_2$ and Y are the same as that in Section 2.3 but replacing

$$\{(U_i,\ Z_{i1}^{\mathrm{T}},\ Z_{i2}^{\mathrm{T}},\ y_i):i=1,\cdots,n\},$$

by

$$\{(t_{ij},\ Z_{i1}^{\mathrm{T}}(t_{ij}),\ Z_{i2}^{\mathrm{T}}(t_{ij}),\ y_i(t_{ij})):j=1,\cdots,\ J_i,\ i=1,\cdots,\ n\},$$

and W is a weight matrix. When estimators $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}(t_{ij})$ of $\boldsymbol{\theta}$ and $\sigma(t_{ij})$ are available, it is

$$W=\mathrm{diag}\left(\widehat{\sum}_1^{-1},\cdots,\widehat{\sum}_n^{-1}\right),\ \widehat{\sum}_i=\widehat{V}_iC_i(\widehat{\theta})\widehat{V}_i,\ \widehat{V}_i=\mathrm{diag}(\widehat{\sigma}(t_{i1}),\cdots,\widehat{\sigma}(t_{iJ_i})),$$

where $C_i(\boldsymbol{\theta})$ is the correlation matrix of $\boldsymbol{\varepsilon}_i$.

After obtaining the estimator $\hat{\mathbf{a}}_2$ of $\mathbf{a}_2$, we substitute $\hat{\mathbf{a}}_2$ for $\mathbf{a}_2$ in model (4.3) and apply the estimation in Section 2.1.1 to get the estimator $\hat{\mathbf{a}}_1(\cdot)$ of $\mathbf{a}_1(\cdot)$. Let

$$r_{ij}=y_i(t_{ij})-Z_{i1}^{\mathrm{T}}(t_{ij})\widehat{\mathbf{a}}_1(t_{ij})-Z_{i2}^{\mathrm{T}}(t_{ij})\widehat{\mathbf{a}}_2,\ \text{and}\ \mathbf{r}_i=(r_{i1},\cdots,r_{iJ_i})^{\mathrm{T}}.$$

A natural estimator of $\sigma^2(t)$ is the kernel estimator

$$\widehat{\sigma}^2(t)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{J_i}r_{ij}^2K_h(t-t_{ij})}{\sum_{i=1}^{n}\sum_{j=1}^{J_i}K_h(t-t_{ij})}.$$

Based on $\mathbf{r}_i$, we can estimate $\boldsymbol{\theta}$ by minimizing the quasi-likelihood function

$$\sum_{i=1}^{n}\left\{\log|C_i(\theta)|+\mathbf{r}_i^{\mathrm{T}}\widehat{V}_i^{-1}C_i^{-1}(\theta)\widehat{V}_i^{-1}\mathbf{r}_i\right\}$$

with respect to $\boldsymbol{\theta}$, and the minimizer is the estimator of $\boldsymbol{\theta}$. We name this estimator quasi-likelihood estimator.

The quasi-likelihood estimator is a good estimator when the correlation structure is correctly specified. However, when the correlation structure is misspecified, the quasi-likelihood estimator may incur a larger bias. Fan *et al.*(2007) proposed another more robust estimator which is based on minimizing the generalized variance of $\mathbf{a}_2$. Explicitly, the estimator of $\boldsymbol{\theta}$ based on the generalized variance method is the minimizer of determinant of the covariance matrix of $\mathbf{a}_2$

$$|\mathbf{D}\mathrm{diag}\left(\widehat{V}_1C_1(\theta)\widehat{V}_1,\cdots,\widehat{V}_nC_n(\theta)\widehat{V}_n\right)\mathbf{D}^{\mathrm{T}}|,$$

where

$$\mathbf{D}=\left\{\mathbf{Z}_2^{\mathrm{T}}(I_n-\mathbf{S})^{\mathrm{T}}W(I_n-\mathbf{S})\mathbf{Z}_2\right\}^{-1}\mathbf{Z}_2^{\mathrm{T}}(I_n-\mathbf{S})^{\mathrm{T}}W.$$

Their philosophy is to improve the estimating parametric component $\mathbf{a}_2$ even when the semiparametric model on the covariance structure is wrong, as long as the model includes the current working covariance structure as a specific case. That motivates the aforementioned optimization criterion. Fan and Wu (2007) investigated the asymptotic properties of the modeling parameter $\boldsymbol{\theta}$ under the weaker conditions that $\mathbf{a}_1(\cdot)$ can be rough and introduced a difference-based method to reduce the bias in estimating $\mathbf{a}_1(\cdot)$.

5 Survival analysis

Survival analysis is an important subject in statistics. It has been widely used in medical science, economics, finance, social science, among others. The most popular model in survival analysis is the Cox model proposed by Cox (1972), which assumes the hazard function $h(t|X)$ of the survival time T is the following proportional hazard function

$$h(t|X)=h_0(t)\exp\left\{X^{\mathrm{T}}\beta\right\}, \tag{5.1}$$

where X is a p dimensional covariate, and $h_0(t)$ is the baseline hazard function. Cox (1972) also proposed the partial likelihood estimation to estimate $\boldsymbol{\beta}$.

Whilst the Cox model is very successful in many applications, it doesn't address any dynamical feature which may exist in the data set. Zhang and Steele (2004) studied the data set about the contraceptive use in the Bangladesh, and found a very strong dynamical pattern with the data set. From socioeconomic point of view, this kind of dynamical pattern is very important as it may reveal how the society and political system change with time. To address the dynamical feature, Zhang and Steele (2004) proposed a semiparametric multilevel survival model. On the individual level, their model can be viewed as a special case of the varying coefficient proportional hazard function models

$$h(t|X,U)=h_0(t)\exp\left\{X^{\mathrm{T}}\mathbf{a}(U)\right\}, \tag{5.2}$$

where U is a scalar covariate. Fan *et al.*(2006) systematically studied models (5.2). They proposed the local partial likelihood estimation to estimate $\mathbf{a}(\cdot)$, and derived the asymptotic normality of their estimators.

The local partial likelihood estimation is outlined as follows: Suppose we have a sample $(U_i, X_i^{\mathrm{T}}, y_i, \delta_i)$, $i = 1, \cdots, n$. $y_i = \min(T_i, C_i)$, $\delta_i = I(T_i > C_i)$, T_i and C_i are respectively the survival time and censoring time of the ith sample member. The censoring mechanism is assumed to be noninformative. Further, denote the distinct event times by $y_{(1)} < \cdots < y_{(L)}$ and the number of events at time $y_{(\ell)}$ by d_ℓ. Denote the set of indices for the individuals at risk up to time $y_{(\ell)}$ by R_ℓ, and the set of indices for the events at $y_{(\ell)}$ by $\mathcal{D}_\ell$. For any given u, the local partial likelihood estimator of $\mathbf{a}(u)$ is the part corresponding to $\mathbf{a}$ of the maximizer of the following local partial log-likelihood function

$$\sum_{\ell=1}^{L}\left(\sum_{j\in\mathcal{D}_\ell}K_h(U_j-u)\left[X_j^{\mathrm{T}}\{\mathbf{a}+\mathbf{b}(U_j-u)\}-\log\left(\sum_{k\in R_\ell}\exp\left[X_k^{\mathrm{T}}\{\mathbf{a}+\mathbf{b}(U_j-u)\}\right]K_h(U_k-u)\right)\right]\right).$$

Fan *et al.*(2006) also discussed the estimation for the bias and variance of the local partial likelihood estimators, as well as the variable selection issue.

Cai *et al.*(2007a) successfully extended the local partial likelihood estimation to multivariate survival data with partially linear hazard regression. They proposed a profile pseudo-partial likelihood estimation. An iterative algorithm was developed to implement the estimation. They also established the asymptotic normality of their estimators. The estimation for standard error as well as hypothesis test for the parametric component are also discussed.

Cai *et al.*(2007b) investigated the semivarying coefficient hazard regression models for multivariate survival data. Cai *et al.*(2007c) studied the marginal varying coefficient hazard models for multivariate survival data. The B-Splines based estimation for model (5.2) was established by Nan *et al.*(2005).

Tian *et al.*(2005) studied a slightly different varying coefficient proportional hazard function models

$$h(t|X)=h_0(t)\exp\left\{X^{\mathrm{T}}\mathbf{a}(t)\right\}. \tag{5.3}$$

The difference between (5.2) and (5.3) is the U with $\mathbf{a}(U)$ in (5.2) is observable, however, the t with $\mathbf{a}(t)$ in (5.3) is the survival time which may be censored. The model (5.3) can still be estimated by local partial likelihood approach, see Tian *et al.*(2005). Pointwise confidence intervals and confidence bands of $\mathbf{a}(\cdot)$ in model (5.3) are also discussed by Tian *et al.*(2005).

6 Nonlinear time series

Varying coefficient models have been elegantly applied to modeling and predicting time series data (Nicholls and Quinn 1982; Chen and Tsay, 1993; Cai, *et al.*, 2000; Huang and Shen, 2004). They are natural extensions of the threshold autoregression models, extensively discussed in Tong (1990). Let $\{X_t\}$ be a given time series. The varying coefficient model is of the form,

$$X_t=a_0(X_{t-p})+a_1(X_{t-p})X_{t-1}+\cdots+a_k(X_{t-p})X_{t-k}+\varepsilon_t, \tag{6.1}$$

for some given lags k and p. The geometric ergodicity of this model was studied by Chen and Tsay (1993).

The local linear method applies readily to this autoregressive setting. The coefficient functions can be fitted using the local linear technique in Section 2.1.1 by setting

$$Y_i=X_i,\ U_i=X_{i-p},\ X_{i0}=1,\ X_{i1}=X_{i-1},\cdots,\ X_{ik}=X_{i-k},\ i=\max(p+1,k+1),\cdots,n,$$

if the observed time series is $X_1, \cdots, X_n$. The joint asymptotic normality of such an estimator has been studied in Cai *et al.*(2000). They proposed a method for bandwidth and a generalized pseudo-likelihood test for testing the autoregressive models and thresholded models. The method has been successfully applied by Hong and Lee (2003) for inference and forecast of exchange rates.

The varying variable U is taken to be X_{t-p} in (6.1). Fan *et al.*(2003) allows the linear index $\beta_1X_{t-1}+\cdots+\beta_kX_{t-k}$ to be the variable U. See (2.8) for additional details. In particular, it allows U to be allow of the lag variables, not just a given X_{t-p}.

7 Time-varying diffusion models

Diffusion models are frequently used to describe the dynamics of stock prices and interest rates. Let X_t be the log return of stock price or interest rate at time t. The one factor model postulates that X_t satisfies a time-dependent continuous-time stochastic differential equation:

$$dX_t=\mu(t,X_t)\,dt+\sigma(t,X_t)\,dW_t. \tag{7.1}$$

Here W_t denotes the standard Brownian motion and the bivariate functions $\mu(t, X_t)$ and $\sigma(t, X_t)$ are called the instaneous return and volatility of the process $\{X_t\}$ respectively. See, for example, Duffie (1996) and Hull (2003). However, one can not estimate the bivariate functions μ and σ nonparametrically, as we only observe a trajectory (t, X_t) on the bivariate space, which is not dense. Therefore, further restrictions are needed.

One specification is the time-homogenous diffusion model:

$$dX_t=\mu(X_t)\,dt+\sigma(X_t)\,dW_t. \tag{7.2}$$

The nonparametric model has been thoroughly studied by Stanton (1997), Fan and Yao (1998), Chapman and Pearson (2000), and Fan and Zhang (2003), among others. It includes many famous families of parametric models popularly used in the finance literature such as the geometric Brownian motion for stock prices, and interest rate models of Vasicek (1977), Cox, Ingersoll and Ross (CIR) (1985), Chan Karolyi, Longstaff and Sanders (CKLS) (1992), among others.

Economic conditions change from time to time. Thus, it is reasonable to expect that the instantaneous expected return and volatility depend on both time and price level for a given state variable. To take this and estimability into consideration, Fan *et al.*(2003) proposed the following time-varying coefficient model:

$$dX_t=\{\alpha_0(t)+\alpha_1(t)X_t\}dt+\beta_0(t)X_t^{\beta_1(t)}dW_t. \tag{7.3}$$

This is an extension of the CKLS model when all varying coefficients are indeed constant. It is also an extension of the famous CIR model with

$$\alpha_0(t)=\alpha_0,\ \alpha_1(t)=\alpha_1,\ \beta_0(t)=\beta_0,\ \beta_1(t)=1/2$$

for modeling the short-term interest rate. Geometric Brown motion corresponds to

$$\alpha_0(t)=0,\ \alpha_1(t)=\mu,\ \beta_0(t)=\sigma,\ \beta_1=1,$$

in (7.3).

Suppose that the process is observed at discrete time points with the data $\{X_{t_i}, i = 1, \ldots, n+1\}$. Denote by

$$Y_{t_i}=X_{t_{i+1}}-X_{t_i},\ Z_{t_i}=W_{t_{i+1}}-W_{t_i},\ \text{and}\ \Delta_i=t_{i+1}-t_i.$$

According to the independent increment property of the Brownian motion, $\{Z_{t_i}\}$ are independent and normally distributed with mean zero and variance Δ_i. Thus, the discretized version of (7.3) can be expressed as

$$Y_{t_i}\approx\{\alpha_0(t_i)+\alpha_1(t_i)X_{t_i}\}\Delta_i+\beta_0(t_i)X_{t_i}^{\beta_1(t_i)}\sqrt{\Delta_i}\varepsilon_{t_i},\ i=1,\ldots,n, \tag{7.4}$$

where $\{\varepsilon_{t_i}\}_{i=1}^n$ are independent and have a standard normal distribution. This is indeed a vary coefficient model in both the conditional mean and conditional variance.

Fan *et al.*(2003) employed the local constant approach to estimate the coefficients $\alpha_0(t)$ and $\alpha_1(t)$ in a similar manner to Section 2.1.1, i.e. minimizing with respect to a and b

$$\sum_{i=1}^{n}\left[\frac{Y_{t_i}}{\Delta_i} - a - bX_{t_i}\right]^2 K_h(t_i - t_0)$$

for each given t_0, resulting the estimators $\hat{\alpha_0}(t) = \hat{a}$ and $\hat{\alpha_1}(t_0) = \hat{b}$, with $\hat{a}$ and $\hat{b}$ being the minimizer of the above local linear-squares problem. The reason that the local constant instead of local linear technique is used is to avoid small arbitrary linear trend, created by the local linear fit with a large bandwidth. After the time-varying coefficients $\alpha_0(t)$ and $\alpha_1(t)$ were estimated, they employed the pseudo-likelihood method to estimate $\beta_0(t)$ and $\beta_1(t)$.

Let $\widehat{E}_t = \{Y_t - (\widehat{\alpha}_0(t) + \widehat{\alpha}_1(t)X_t)\Delta_t\}/\sqrt{\Delta_t}$. Then, by (7.4), we have

$$\widehat{E}_t \approx \beta_0(t) X_t^{\beta_1(t)} \varepsilon_t. \tag{7.5}$$

At each given point t_0, the following local pseudo-likelihood, which is the local normal-likelihood if (7.5) holds exactly,

$$\ell(\beta_0, \beta_1; t_0) = -\frac{1}{2}\sum_{i=1}^{n} K_h(t_i - t_0)\left(\log(\beta_0^2 X_{t_i}^{2\beta_1}) + \frac{\widehat{E}_{t_i}^2}{\beta_0^2 X_{t_i}^{2\beta_1}}\right)$$

is maximized, yielding $\hat{\beta}_0(t_0) = \hat{\beta}_0$ and $\hat{\beta}_1(t_0) = \hat{\beta}_1$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the maximizer of the above local maximum likelihood estimator. Note that given β_1, $\ell(\beta_0, \beta_1; t_0)$ can be maximized explicitly with respect to β_0 and therefore, the problem reduces to the univariate optimization problem.

Fan *et al.*(2003) suggested using one-sided kernel so that only past data are used. They also proposed a method to select the bandwidth and to construct confidence intervals. The model was also used as the alternative hypothesis for testing the famous time homogeneous model such as the CIR and CKLS models. It was also applied to price zero-coupon bond.

8 Concluding remark

In this paper, we have given a selective overview on the developments on the varying coefficient models. There are a vast of number of papers addressing various types of varying coefficient models in the past ten years. Our citation in this paper is not exhaustive. In addition to the applications to time series, longitudinal data analysis and survival analysis, the varying coefficient models have also seen their applications in other subjects in statistics. For example, Sentürk and Müller (2005) applied the varying coefficient models in covariate adjusted correlation analysis.

We only focus on the major developments on the standard varying coefficient models and their extensions in time series, longitudinal data analysis and survival analysis. Our emphasis is placed on the methodological side. We have not cited the papers with main contributions on applied side. Undoubtedly, varying coefficient models have seen their broad and exciting applications in many scientific areas in the last ten years. Examples include that Ferguson *et al.*(2007) applied the varying coefficient models to explore the complex ecological system at Loch Leven, and obtained some insight into the combined effects of climate change and eutrophication on water quality; Kauermann *et al.*(2005) used the varying coefficient models to analyse the survival of 1123 newly founded firms in the state of Bavaria, Germany, and investigate the time varying effects of risk factors. The varying coefficient models are becoming more and more attractive to both applied and methodological statisticians. They are being more and more frequently used in many scientific areas to explore the dynamic feature.

References

Ahmad I, Leelahanon S, Li Q. Efficient estimation of a semiparametric partially linear varying coefficient model. The Annals of Statistics 2005;33:258–283.

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

Brumback B, Rice JA. Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). J Amer Statist Assoc 1998;93:961–994.

Cai J, Fan J, Jiang J, Zhou H. Partially linear hazard regression for multivariate survival data. Journal of American Statistical Association 2007a;102:538–551.

Cai J, Fan J, Jiang J, Zhou H. Partially linear hazard regression with varying-coefficients for multivariate survival data. Journal Royal Statistical Society, B. 2007bto appear

Cai J, Fan J, Zhou H, Zhou Y. Marginal hazard models with varying-coefficients for multivariate failure time data. The Annals of Statistics 2007c;35:324–354.

Cai Z, Fan J, Li R. Efficient estimation and inferences for varying-coefficient models. Jour Ameri Statist Assoc 2000;95:888–902.

Cai Z, Fan J, Yao Q. Functional-coefficient regression models for nonlinear time series. Journal of American Statistical Association 2000;95:941–956.

Carroll RJ, Fan J, Gijbels I, Wand MP. Generalized partially linear single-index models. J Amer Statist Assoc 1997;92:477–489.

Carroll RJ, Ruppert D, Welsh AH. Local estimating equations. Jour Ameri Statist Assoc 1998;93:214–227.

Chan KC, Karolyi AG, Longstaff FA, Sanders AB. An empirical comparison of alternative models of the short-term interest rate. Journal of Finance 1992;47:1209–1227.

Chapman DA, Pearson ND. Is the short rate drift actually nonlinear? Journal of Finance 2000;55:355–388.

Chen R, Tsay RJ. Functional-coefficient autoregressive models. J Amer Statist Assoc 1993;88:298–308.

Cheng M-Y, Zhang W. Statistical estimation in generalized multiparameter likelihood models. manuscript. 2007

Chiang CT, Rice JA, Wu CO. Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. Jour Amer Statist Assoc 2001;96:605–619.

Cleveland, WS.; Grosse, E.; Shyu, WM. Local regression models. In: Chambers, JM.; Hastie, TJ., editors. Statistical Models in S. Wadsworth & Brooks; Pacific Grove: 1991. p. 309-376.

Chapman DA, Pearson ND. Is the short rate drift actually nonlinear? Journal of Finance 2000;55:355–388.

Cox DR. Regression models and life tables (with discussion). J Roy Statist Soc Ser B 1972;34:187–220.

Cox JC, Ingersoll JE, Ross SA. A theory of the term structure of interest rates. Econometrica 1985;53:385–467.

Diggle, PJ.; Heagerty, P.; Liang, KY.; Zeger, SL. Analysis of Longitudinal Data. Oxford University Press; 2002.

Duffie, D. Dynamic Asset Pricing Theory. 2. Princeton University Press; Princeton, N.J: 1996.

Fan J, Gijbels I. Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. J Royal Statist Soc B 1995;57:371–394.

Fan, J.; Gijbels, I. Local Polynomial Modelling and Its Applications. Chapman and Hall; London: 1996.

Fan J, Farmen M, Gijbels I. Local maximum likelihood estimation and inference. Journal of Royal Statistical Society B 1998;60:591–608.

Fan J, Jiang J, Zhang C, Zhou Z. Time-dependent diffusion models for term structure dynamics and the stock price volatility. Statistica Sinica 2003;13:965–992.

Fan J, Härdle W, Mammen E. Direct estimation of additive and linear components for high dimensional data. The Annals of Statistics 1998;26:943–971.

Fan J, Huang T. Profile Likelihood Inferences on semiparametric varying-coefficient partially linear models. Bernoulli 2005;11:1031–1057.

Fan J, Huang T, Li RZ. Analysis of longitudinal data with semiparametric estimation of covariance function. Journal of American Statistical Association 2007;35:632–641.

Fan J, Li R. New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. Jour Ameri Statist Assoc 2004;99:710–723.

Fan J, Peng H. On non-concave penalized likelihood with diverging number of parameters. The Annals of Statistics 2004;32:928–961.

Fan J, Wu Y. Semiparametric estimation of covariance matrices for longitudinal data. Manuscript. 2007

Fan J, Yao Q. Efficient estimation of conditional variance functions in stochastic regression. Biometrika 1998;85:645–660.

Fan J, Yao Q, Cai Z. Adaptive varying-coefficient linear models. Journal of Royal Statistical Society B 2003;65:57–80.

Fan J, Zhang CM, Zhang J. Generalized likelihood ratio statistics and Wilks phenomenon. The Annals of Statistics 2001;29:153–193.

Fan J, Zhang JT. Functional linear models for longitudinal data. Jour Roy Statist Soc B 2000;62:303–322.

Fan J, Zhang C. A reexamination of diffusion estimations with applications to financial model validation. Journal of American Statistical Association 2003;98:118–134.

Fan J, Zhang W. Statistical estimation in varying coefficient models. Ann Statist 1999;27:1491–1518.

Fan J, Zhang W. Simultaneous confidence bands and hypothesis testing in varying-Coefficient models. Scand J Statist 2000;27:715–731.

Ferguson CA, Bowman AW, Scott EM. Model comparison for a complex ecological system. Jour Roy Statist Soc A 170:691–711.

Friedman JH. Multivariate adaptive regression splines (with discussion). Ann Statist 1991;19:1–141.

Green, PJ.; Silverman, BW. Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach. Chapman and Hall; London: 1994.

Gu C, Wahba G. Smoothing spline ANOVA with component-wise Bayesian “confidence intervals”. J Comput Graph Statist 1993;2:97–117.

Hand, D.; Crowder, M. Practical Longitudinal Data Analysis. Chapman and Hall; London: 1996.

Härdle W, Stoker TM. Investigating smooth multiple regression by the method of average derivatives. J Amer Statist Assoc 1989;84:986–995.

Hart JD, Wehrly TE. Consistency of cross-validation when the data are curves. Stochastic Process and their Applications 1993;45:351–361.

Hastie, TJ.; Tibshirani, RJ. Generalized Additive Models. London: Chapman and Hall; 1990.

Hastie TJ, Tibshirani RJ. Varying-coefficient models. Jour Roy Statist Soc B 1993;55:757–796.

Heckman J, Ichimura H, Smith J, Todd P. Characterizing Selection Bias Using Experimental Data. Econometrica 1998;66:1017–1098.

Hogan JW, Lin X, Herman B. Mixtures of varying coefficient models for longitudinal data with discrete or continuous nonignorable dropout. Biometrics 2004;60:854–864. [PubMed: 15606405]

Hong Y, Lee TH. Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. Review of Economics and Statistics 2003;85:1048–1062.

Hoover DR, Rice JA, Wu CO, Yang LP. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika 1998;85:809–822.

Huang JZ, Wu CO, Zhou L. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. Biometrika 2002;89:111–128.

Huang JZ, Shen H. Functional coefficient regression models for nonlinear time series: A polynomial spline approach. Scandinavian Journal of Statistics 2004;31:515–534.

Huang JZ, Wu CO, Zhou L. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. Statistica Sinica 2004;14:763–788.

Huber PJ. Projection pursuit (with discussion). Ann Statist 1985;13:435–525.

Hull, J. Options, Futures, and Other Derivatives. 15. Prentice Hall; Upper Saddle Review, New Jersey: 2003.

Kauermann G, Tutz G. On model diagnostics using varying coefficient models. Biometrika 1999;86:119–128.

Kauermann G, Tutz G, Brüderl J. The survival of newly founded firms: a case-study into varying-coefficient models. Jour Roy Statist Soc A 2005;168:145–158.

Li K-C. Sliced inverse regression for dimension reduction (with discussion). J Amer Statist Assoc 1991;86:316–342.

Li R, Liang H. Variable selection in semiparametric regression modeling. Annals of Statistics. 2007To appear

Lin DY, Ying Z. Semiparametric and nonparametric regression analysis of longitudinal data (with discussions). Jour Amer Statist Assoc 2001;96:103–126.

Lin X, Carroll RJ. Semiparametric regression for clustered data using generalized estimating equations. Jour Amer Statist Assoc 2001b;96:1045–1056.

Nan B, Lin X, Lisabeth LD, Harlow S. A varying-coefficient Cox model for the effect of age at a marker event on age at menopause. Biometrics 2005;61:576–583. [PubMed: 16011707]

Nicholls, DF.; Quinn, BG. Random Coefficient Autoregressive Models: An Introduction. Springer-Verlag; New York: 1982.

Qu A, Li R. Quadratic inference functions for varying coefficient models with longitudinal data. Biometrics 2006;62:379–391. [PubMed: 16918902]

Ramsay, JO.; Silverman, BW. The Analysis of Functional Data. Springer-Verlag; Berlin: 1997.

Rice JA, Silverman BW. Estimating the mean and covariance structure nonparametrically when the data are curves. Jour Roy Statist Soc B 1991;53:233–243.

Sentürk D, Müller HG. Covariate adjusted correlation Analysis via varying coefficient models. Scandinavian Journal of Statistics 2005;32:365–383.

Stanton R. A nonparametric models of term structure dynamics and the market price of interest rate risk. Journal of Finance 1997;52:1973–2002.

Stone CJ, Hansen M, Kooperberg C, Truong YK. Polynomial splines and their tensor products in extended linear modeling. Ann Statist 1997;25:1371–1470.

Sun Y, Zhang W, Tong H. Estimation of the covariance matrix of random effects in longitudinal studies. Ann Statist 2007;35:2795–2814.

Tian L, Zucker D, Wei LJ. On the Cox model with time-varying regression coefficients. Jour Ameri Statist Assoc 2005;100:172–183.

Tong, H. Non-Linear Time Series: A Dynamical System Approach. Oxford University Press; Oxford: 1990.

Wahba, G. Partial spline models for semiparametric estimation of functions of several variables. Statistical Analysis of Time Series, Proceedings of the Japan U.S. Joint Seminar; Tokyo. Tokyo: Institute of Statistical Mathematics; 1984. p. 319-329.

Wang N. Marginal nonparametric kernel regression accounting for within-subject correlation. Biometrika 2003;90:43–52.

Wang N, Carroll RJ, Lin X. Efficient semiparametric marginal estimation for longitudinal/clustered data. Jour Ameri Statist Assoc 2005;100:147–157.

Wu CO, Chiang CT, Hoover DR. Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. Jour Ameri Statist Assoc 1998;93:1388–1402.

Wu CO, Chiang CT. Kernel smoothing on varying coefficient models with longitudinal dependent variable. Statist Sinica 2000;10:433–456.

Vasicek O. An equilibrium characterization of the term structure. Journal of Financial Economics 1977;5:177–188.

Xia Y, Li WK. On the estimation and testing of functional-Coefficient linear models. Statistica Sinica 1999;9:735–758.

Xia Y, Zhang W, Tong H. Efficient estimation for semivarying-coefficient models. Biometrika 2004;91:661–681.

Zeger SL, Diggle PJ. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. Biometrics 1994;50:689–99. [PubMed: 7981395]

Zhang W, Lee SY. Variable bandwidth selection in varying-coefficient models. Journal of Multivariate Analysis 2000;74:116–134.

Zhang W, Lee SY, Song X. Local polynomial fitting in semivarying coefficient models. Jour Multivar Anal 2002;82:166–188.

Zhang W, Steele F. A semiparametric multilevel survival model. Journal of the Royal Statistical Society, Series C 2004;53:387–404.

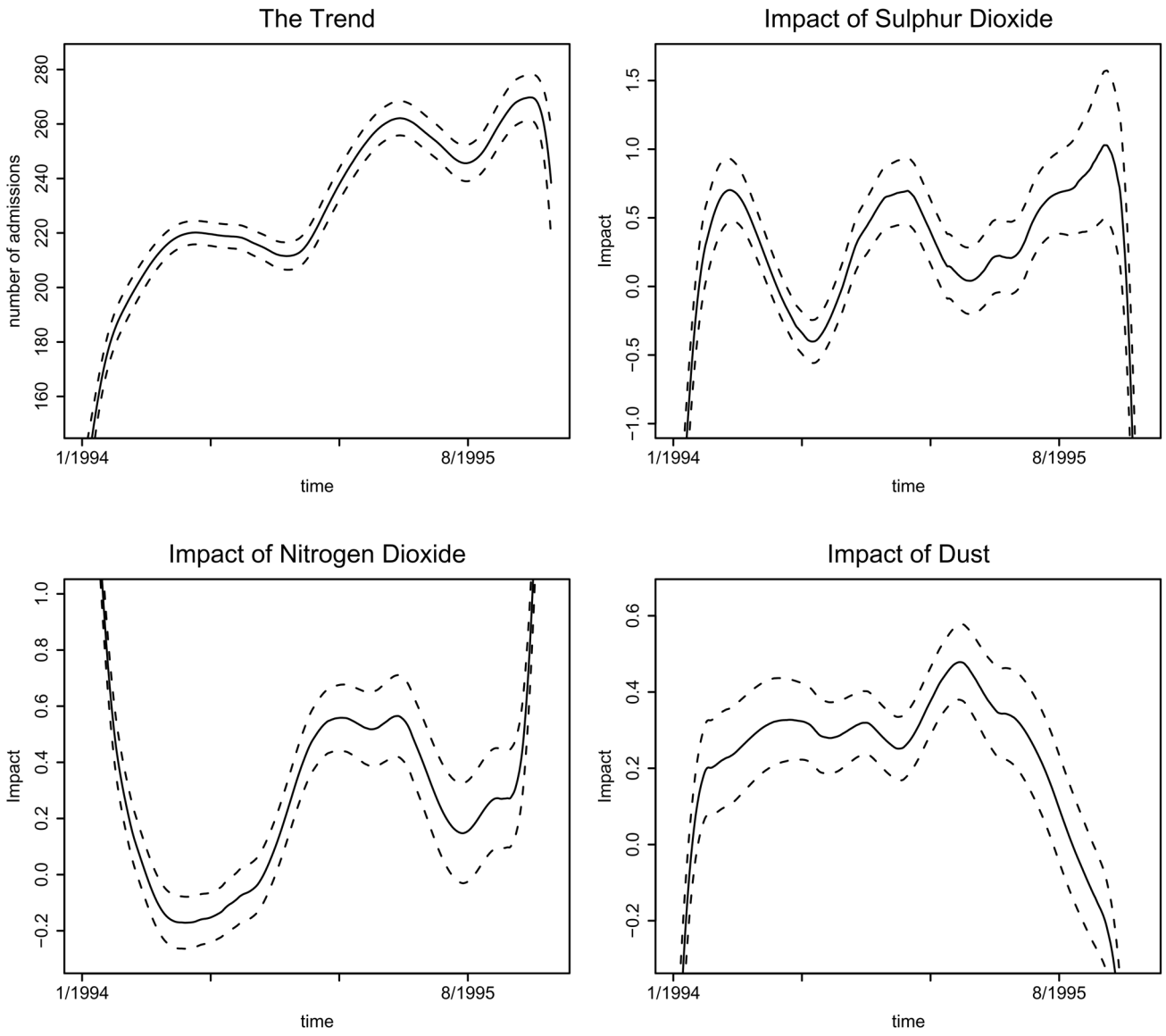

Figure 1.
The impacts of Sulphur Dioxide, Nitrogen Dioxide and Dust on the number of daily total hospital admissions for circulationary and respirationary problems