CHAPTER 1

AN OVERVIEW ON NONPARAMETRIC AND SEMIPARAMETRIC TECHNIQUES FOR LONGITUDINAL DATA

Jianqing Fan[a] and Runze Li[b]

[a] *Department of Operation Research and Financial Engineering*
Princeton University, Princeton, NJ 08544
jqfan@Princeton.EDU

[b] *Department of Statistics and the Methodology Center*
Pennsylvania State University
University Park, PA 16802-2111
rli@stat.psu.edu

In the last two decades, there are considerable literature on the topic of longitudinal data analysis. In particular, many authors have made much effort on developing diverse nonparametric and semiparametric models, along with their inference procedures, for longitudinal data. This chapter presents a review on recent development on this topic.

1. Introduction

Longitudinal data are often highly unbalanced because data were collected at irregular and possibly subject-specific time points. It is difficult to directly apply traditional multivariate regression techniques for analyzing such highly unbalanced collected data. This has led biostatisticians and statisticians to develop various modeling procedures for longitudinal data.

Parametric regression models have been extended to longitudinal data analysis (Diggle, *et al.* 2002). They are very useful for analyzing longitudinal data and for providing a parsimonious description of the relationship between the response variable and its covariates. However, the parametric assumption likely introduce modeling biases. To relax the assumptions on parametric forms, various nonparametric models have been proposed for longitudinal data analysis. Earlier works on nonparametric regression analysis for longitudinal data were summarized in Müller (1988). Kernel regression was applied to repeated measurements data with continuous re-

sponses in Hart and Wehrly (1986), and for data with time-series errors in Altman (1990) and Hart (1991). Time-varying coefficient models and varying-coefficient models for continuous responses were proposed for longitudinal data in Faraway (1997, 1999) Hoover, *et al.* (1998), Wu, *et al.* (1998), Fan and Zhang (2000), Wu and Chiang (2000), Chiang, Rice and Wu (2001), and Huang, Wu and Zhou (2002). Nonparametric regression with a single covariate has also been extended for longitudinal data in the setting of generalized linear models in Lin and Carroll (2000) and Wang (2003). Qu and Li (2005) studied time-varying coefficient models under the generalized linear model framework using the quadratic inference function approach (Qu, Lindsay and Li, 2000). More references will be given in Sections 2 and 4.

Although parametric models may be restrictive for some applications, they are more parsimoneous than nonparametric models may be too flexible to make concise conclusions. Semiparametric models are good compromises and retain nice features of both the parametric and nonparametric models. Thus, various semiparametric models have been extended for longitudinal data. Zeger and Diggle (1994) and Moyeed and Diggle (1994) extended partially linear models for longitudinal data. There are many papers on semiparametric modeling for longitudinal data published in the recent literature (Martinussen and Scheike, 1999, Lin and Carroll, 2001a, 2001b, Lin and Ying, 2001, Fan and Li, 2004, Wang, Carroll and Lin, 2005).

This chapter aims to present a selective overview of recent developments on the topic of nonparametric and semiparametric regression modeling for longitudinal data. A complete, detailed review on this topic is impossible due to the limited space. The rest of this chapter is organized as follows. Section 2 provides a review of nonparametric smoothing procedures for longitudinal data with a single covariate. In Section 3, we summarize recent developments on partially linear models for longitudinal data. A review on time-varying coefficient models and functional linear models is given in Section 4. An illustration is presented in Section 5. Some generalizations of models introduced in Sections 2, 3 and 4 are given in Section 6. We will briefly review the recent developments in estimation of covariance functions for the analysis of longitudinal data.

2. Nonparametric model with a single covariate

Suppose that $\{(x_{ij}, y_{ij}),\ j = 1, \cdots, J_i\}$ is a random sample collected from the i-th subject or cluster, $i = 1, \cdots, n$. In this chapter, we assume that J_i is finite. Denote the conditional mean and variance as $\mu_{ij} = E(y_{ij}|x_{ij})$

and $\sigma_{ij}^2 = \mathrm{Var}(y_{ij}|x_{ij})$, respectively. Here it is assumed that the regression function $E(y_{ij}|x_{ij})$ is a nonparametric smooth function of x_{ij}. For longitudinal data and clustered data, it is known that samples collected within a subject are correlated and samples between subjects are often independent. It has been an interesting topic on how to incorporate within-subject correlation information into estimation of the mean function in the literature. Several statistical models have been proposed in existing works. For example, Ruckstuhl, Welsh and Carroll (2000) and Wu and Zhang (2002) studied the case in which the response variable is normally distributed. Severini and Staniswalis (1994), Lin and Carroll (2000) and Wang (2003) proposed estimation procedures for the marginal mean function under the framework of generalized linear models.

Let us begin with a summary of the work by Lin and Carroll (2000), in which it is assumed that $\sigma_{ij}^2 = \phi_j w_{ij} v(\mu_{ij})$, where ϕ_j is a scale parameter, w_{ij} is a known weight and $v(\cdot)$ is a variance function. Under the framework of generalized linear models, assume that the marginal mean μ_{ij} depends on x_{ij} through a known canonical link function $\mu(\cdot)$,

$$\mu_{ij} = \mu\{\theta(x_{ij})\},$$

where $\theta(\cdot)$ is an unknown smooth function. The local likelihood approach (Fan, Heckman and Wand, 1995) is employed to estimate $\theta(\cdot)$. The use of canonical link guarantees that the associated optimization problem is either a convex minimization or concave maximization.

Here we focus on the local linear method for simplicity of notation. The idea is applicable for local polynomial methods (Lin and Carroll, 2000). To motivate the kernel generalized estimation equations (GEE) approach, let us temporarily assume that $\theta(x)$ is a linear function of x:

$$\theta(x) = \beta_0 + \beta_1 x \equiv \mathbf{g}(x)^T \boldsymbol{\beta},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, and $\mathbf{g}(x) = (1, x)^T$. Let $\mathbf{y}_i = (y_{i1}, \cdots, y_{iJ_i})^T$ and $\boldsymbol{\mu}_i = E(\mathbf{y}_i) = [\mu\{\mathbf{g}(x_{i1})^T\boldsymbol{\beta}\}, \cdots, \mu\{\mathbf{g}(x_{iJ_i})^T\boldsymbol{\beta}\}]^T$. The conventional GEE (Liang and Zeger, 1986) approach estimates $\boldsymbol{\beta}$ by solving the following equations:

$$\sum_{i=1}^{n} \mathbf{G}_i^T \Delta_i \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = 0,$$

where $\mathbf{G}_i = (\mathbf{g}(x_{i1}), \cdots, \mathbf{g}(x_{iJ_i}))^T$, $\Delta_i = \mathrm{diag}\{\mu'\{\mathbf{g}^T(x_{ij})\boldsymbol{\beta}\}\}$, $\mu'(\cdot)$ is the first order derivative of $\mu(\cdot)$, $\mathbf{V}_i = \mathbf{S}_i^{1/2}\mathbf{R}_i(\boldsymbol{\delta})\mathbf{S}_i^{1/2}$, $\mathbf{S}_i =$

diag$[\phi_j w_{ij} v(\mu\{\mathbf{g}^T(x_{ij})\boldsymbol{\beta}\})]$ and $\mathbf{R}_i$ is an invertible working correlation matrix, which possibly depends on a parameter vector $\boldsymbol{\delta}$. The purpose of including the working correlation matrix $\mathbf{R}_i$ is to improve efficiency of $\boldsymbol{\beta}$. When the true correlation matrix is known, one would use it as the working correlation matrix. When there is no knowledge available on the within-subject correlation, the identity matrix is a convenient choice for working correlation matrix. As shown in Liang and Zeger (1986), if the mean function is correctly specified, the choice of $\mathbf{R}_i$ affects only the efficiency, but not the root n consistency of the resulting estimate $\widehat{\boldsymbol{\beta}}$. It has been shown that the resulting estimate is most efficient when $\mathbf{R}_i$ equals the true correlation matrix. The GEE strategy has been extended for a nonparametric smooth function $\theta(x)$ in Lin and Carroll (2000).

The local linear method is to locally, linearly approximate $\theta(z)$ at a neighborhood of x by

$$\theta(z) = \theta(x) + \theta'(x)(z-x) \equiv \boldsymbol{\beta}^T \mathbf{g}(z-x). \tag{2.1}$$

Here we slightly abuse the notation of $\boldsymbol{\beta}$, which should depend on the given point x. In particular, in this local modeling the first component β_1 of $\boldsymbol{\beta}$ represents $\theta(x)$. Let $K(x)$ be a symmetric kernel density function and h be a bandwidth. Denote $K_h(x) = h^{-1}K(x/h)$. As direct extension of the conventional GEE approach, Lin and Carroll (2000) suggested two ways incorporating kernel weight function in generalized estimation equation. The two ways lead to two sets of kernel GEE's for $\boldsymbol{\beta}$:

$$\sum_{i=1}^{n} \mathbf{G}_i^T(x)\Delta_i(x)\mathbf{V}_i^{-1}(x)\mathbf{K}_{ih}(x)\{\mathbf{y}_i - \boldsymbol{\mu}_i(x)\} = 0, \tag{2.2}$$

and

$$\sum_{i=1}^{n} \mathbf{G}_i^T(x)\Delta_i(x)\mathbf{K}_{ih}^{1/2}(x)\mathbf{V}_i^{-1}(x)\mathbf{K}_{ih}^{1/2}(x)\{\mathbf{y}_i - \boldsymbol{\mu}_i(x)\} = 0, \tag{2.3}$$

where $\mathbf{K}_{ih}(x) = \text{diag}\{K_h(x_{ij}-x)\}$, and $\{\boldsymbol{\mu}_i(x), \Delta_i(x), \mathbf{V}_i(x), \mathbf{S}_i(x)\}$ are the same as those in the conventional GEE except that they are evaluated at $\mu_{ij} = \mu\{\mathbf{g}(x_{ij}-x)^T\boldsymbol{\beta}\}$. Solving equation (2.2) or (2.3) yields an estimate of $\widehat{\boldsymbol{\beta}}$. Having estimated $\boldsymbol{\beta}$ at x, let $\widehat{\theta}(x) = \widehat{\beta}_1$, the first component of $\boldsymbol{\beta}$. The standard error for $\widehat{\theta}(x)$ can be estimated by a sandwich formula, which is a conventional technique in GEE.

Severini and Staniswalis (1994) suggested letting $\mathbf{R}_i$ be an estimator of the actual correlation matrix. Lin and Carroll (2000) showed that it is generally the best strategy to ignore entirely the correlation structure

within each subject/cluster and to instead pretend that all observations are independent. This implies that the behavior of kernel GEE is quite different from that of the parametric GEE. The intuition is that in a local neighborhood around x, there are unlikely to have more than one data point contributed from a subject so that the effective data points used in fitting (2.2) and (2.3) are nearly independent.

Lin and Carroll (2001b) extended the kernel GEE to generalized partially linear model for longitudinal data. Their works inspired other authors to study how to incorporate correlation information into nonparametric regression with longitudinal data. Peterson, Zhao and Eapen (2003) proposed a simple extension of the local polynomial regression smoother that retains the asymptotic properties of the working independence estimator, while typically reducing both the conditional bias and variance for practical sample sizes. Xiao, *et al.* (2003) proposed a modification of local polynomial time series regression estimators that improves efficiency when innovation process is autocorrelated. Wu and Zhang (2002) considered local polynomial mixed effects (LLME) models for longitudinal data with continuous response. They proposed an estimation procedure for the LLME models using local polynomial regression. Their procedure incorporates correlation information by introducing a subjectwise random intercept function. They showed that the asymptotic bias and variance essentially the same as those of the kernel GEE. For finite sample performance, they empirically demonstrated that their procedure is more efficient than the kernel GEE approach.

Wang (2003) illustrated how the kernel GEE methods account within-subject correlation and found that the kernel GEE method uses kernel weights to control biases. Unfortunately, the kernel weight eliminates biases but also eliminates input from correlated elements in the same subject. It is challenging to control bias and to reduce the variation simultaneously. Wang (2003) proposed the marginal kernel method for accomplishing both tasks. The idea is closely related to the method of using control variable for variance reduction in the simulation literature. See, for example, Ross (1997). Suppose that we want to evaluate $\mu_f = Ef(X)$ and it is known from the side information that $E\mathbf{g}(X) = 0$ for a vector of functions $\mathbf{g}$. Then, for any given constant vector $\mathbf{a}$, $\mu_f = E\{f(X) - \mathbf{a}^T\mathbf{g}(X)\}$ can be estimated by its sample average from the simulation. The optimal choice of $\mathbf{a}$ is to minimize $\mathrm{Var}\{f(X) - \mathbf{a}^T\mathbf{g}(X)\}$ with respect to $\mathbf{a}$. This results in the optimal choice $\mathbf{a} = \mathrm{Var}(\mathbf{g})^{-1}\mathrm{cov}(\mathbf{g}, f)$.

The innovation of Wang (2003) is to create the side information via a preliminary estimate and to subtract it from the observed data so that

the observed data have approximate mean zero, which plays the same role as $E\mathbf{g}(X) = 0$. Let $\breve{\theta}(x)$ be a consistent estimator of $\theta(x)$. For example, $\breve{\theta}(x)$ might be taken to be the resulting estimate of the kernel GEE with working independent correlation matrix. Let $\mathbf{G}^i_{*j}$ be an $J_i \times 2$ matrix with the first column being e_j and the second column being $e_j(x - x_{ij})/h$, where e_j denotes the indicator vector with jth entry equal to 1. Wang (2003) proposed solving the following kernel-weighted estimation equation with respect to β_0 and β_1:

$$0 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J_i} K_h(x - x_{ij})\mu'\{\beta_0 + \beta_1(x - x_{ij})\}(\mathbf{G}^i_{*j})^T\mathbf{V}_i^{-1}(x)(\mathbf{y}_i - \boldsymbol{\mu}_{*j})$$

where the l-th element of $\boldsymbol{\mu}_{*j}$ is $\mu\{\beta_0 + \beta_1(x - x_{il})/h\}$, when $l = j$, and is $\mu\{\breve{\theta}(x_{ij})\}$, when $l \neq j$. The basic idea behind the equations is as follows: once a data point, (x_{ij}, y_{ij}) say, within a cluster has its x-value within a local neighborhood of x and is used to estimate $\theta(x)$, all data points in that cluster are used. To improve the efficiency, as in the simulation literature, the contributions of all points but point (x_{ij}, y_{ij}) within cluster to the local estimate of $\theta(x)$ are through their residuals $(\mathbf{y}_i - \boldsymbol{\mu}_{*j})$, namely the side information $E(\mathbf{y}_i - \boldsymbol{\mu}_{*j}) \approx 0$ is used. Denote the solution to be $(\widehat{\beta}_0, \widehat{\beta}_1)^T$. Then $\widehat{\theta}(x) = \widehat{\beta}_0$. As shown in Wang (2003), the smallest variance of the $\widehat{\theta}(x)$ is achieved when the true correlation is employed. Asymptotically, the smallest variance is uniformly smaller than that of the most efficient estimate of kernel GEE.

It is well known that for independent data, kernel regression and spline smoothing are asymptotically equivalent for nonparametric model with a single covariate (Silverman, 1984). Welsh *et al.* (2002) shows this is not the case for longitudinal/clustered data. Splines and conventional kernels are different in localness and ability to account for within-cluster correlation. Lin, *et al.* (2004) showed that a smoothing spline estimator is asymptotically equivalent to the marginal kernel method proposed in Wang (2003). They further showed that both the marginal kernel method and the smoothing spline estimator are nonlocal unless working independence is assumed but have asymptotically negligible bias.

Carroll, *et al.* (2004) proposed histospline method for nonparametric regression models for clustered longitudinal data. This provides a simple approach to deal with the difficulty which arises in situations in which aspects of the regression function are known parametrically, or the sampling scheme has significant structure. The histospline technique converts a prob-

lem in the continuum to one that is governed by only a finite number of parameters.

3. Partially linear models

A natural extension of the nonparametric model with a single covariate and the ordinary linear regression model with multiple covariates is the partially linear model, which has been well studied for independent data in the literature (Wahba, 1984, Engle, *et al.* , 1986, Heckman, 1986, Speckman, 1988, Härdle, Liang and Gao, 1999). This accommodates multiple covariates while retaining a nonparametric baseline function. Let y be a response variable and u and $\mathbf{x}$ be covariates. A partially linear model is defined as

$$y = \alpha(u) + \mathbf{x}^T\boldsymbol{\beta} + \varepsilon, \tag{3.1}$$

where $\alpha(\cdot)$ is a nonparametric smooth baseline function, $\boldsymbol{\beta}$ is an unknown regression coefficient vector and ε is a random error with $E(\varepsilon|u, \mathbf{x}) = 0$. For independent data, Heckman (1986) parameterized $\alpha(u)$ by splines and used the smoothing spline approach to estimate both α and $\boldsymbol{\beta}$. Note that

$$E(y|u) = \alpha(u) + E(\mathbf{x}|u)^T\boldsymbol{\beta}.$$

Then

$$y - E(y|u) = \{\mathbf{x} - E(\mathbf{x}|u)\}^T\boldsymbol{\beta} + \varepsilon. \tag{3.2}$$

Speckman (1988) suggested smoothing y and $\mathbf{x}$ over u and substituting the estimates of $E(y|u)$ and $E(\mathbf{x}|u)$ into (3.2). Thus, we can estimate $\boldsymbol{\beta}$ easily. This approach is referred to as a partial residual approach or more generally a profile least squares method. Plugging-in the estimate of $\boldsymbol{\beta}$ into (3.1), we can further estimate $\alpha(u)$ using one-dimensional smoothing techniques. We next present some existing estimation procedures for partially linear models with longitudinal data.

3.1. *Estimation procedures*

Suppose that we have a sample of n subjects. For the i-th subject, the response variable $y_i(t)$ and the covariate vector $\mathbf{x}_i(t)$, are collected at time points $t = t_{i1}, \cdots, t_{iJ_i}$, where J_i is the total number of observations on the i-th subject. A partially linear model for longitudinal data has the following form:

$$y_i(t_{ij}) = \alpha(t_{ij}) + \boldsymbol{\beta}^T\mathbf{x}_i(t_{ij}) + \varepsilon_i(t_{ij}) \tag{3.3}$$

for $i = 1, \cdots, n$, and $j = 1, \cdots, J_i$. Zeger and Diggle (1994) suggested using a backfitting algorithm to find an estimate for $\alpha(u)$ and $\boldsymbol{\beta}$. Specifically, starting with an initial value of $\boldsymbol{\beta}$, denoted by $\boldsymbol{\beta}^{(0)}$, we smooth the residual $y_i(t_{ij}) - \mathbf{x}_i^T(t_{ij})\boldsymbol{\beta}^{(0)}$ over t_{ij} to estimate $\alpha(\cdot)$. Having an estimate for $\alpha(\cdot)$, denoted by $\widehat{\alpha}(\cdot)$, we conduct linear regression of $y_i(t_{ij}) - \widehat{\alpha}(t_{ij})$ on $\mathbf{x}_i(t_{ij})$. Iterate this procedure until it converges. This is basically the same as the Gauss-Seidal algorithm to compute the profile least squares estimate. Moyeed and Diggle (1994) proposed an improved version of the backfitting algorithm based on a partial residual approach.

Lin and Ying (2001) introduced the counting process technique to the estimation scheme. The time points where the observations on the i-th subject are made are characterized by the counting process: $N_i(t) \equiv \sum_{j=1}^{J_i} I(t_{ij} \leq t)$, where $I(\cdot)$ is the indicator function. Both $y(t)$ and time-varying covariates $\mathbf{x}(t)$ were observed at the jump points of $N_i(t)$. The observation times are regarded as realizations from an arbitrary counting process that is censored at the end of follow-up. Specifically, $N_i(t) = N_i^*(t \wedge c_i)$, where $N_i^*(t)$ is a counting process in discrete or continuous time, c_i is the follow-up or censoring time, and $a \wedge b = \min(a, b)$. The censoring time c_i is allowed to depend on the vector of covariates $\mathbf{x}_i(\cdot)$ in an arbitrary manner. It is assumed that the censoring mechanism is noninformative in the sense that $E\{y_i(t)|\mathbf{x}_i(t), c_i \geq t\} = E\{y_i(t)|\mathbf{x}_i(t)\}$. Lin and Ying (2001) proposed minimizing the following least squares function with counting process notation,

$$\sum_{i=1}^{n} \int_0^{+\infty} w(t)\{y_i(t) - \alpha(t) - \boldsymbol{\beta}^T \mathbf{x}_i(t)\}^2 \, dN_i(t), \tag{3.4}$$

where $w(t)$ is a possibly data-dependent weight function.

Lin and Ying (2001) allow that the potential observation times to depend on the covariates and assume that

$$E\{dN_i^*(t)|\mathbf{x}_i(t), y_i(t), c_i \geq t\} = \exp\{\boldsymbol{\gamma}'\mathbf{x}_i(t)\} \, d\Lambda(t), \quad i = 1, \cdots, n, \tag{3.5}$$

where $\boldsymbol{\gamma}$ is a vector of unknown parameters and $\Lambda(\cdot)$ is an arbitrary nondecreasing function. Denote

$$\bar{\mathbf{x}}(t, \boldsymbol{\gamma}) = \frac{\sum_{i=1}^n \xi_i(t) \exp\{\boldsymbol{\gamma}^T \mathbf{x}_i(t)\} \mathbf{x}_i(t)}{\sum_{i=1}^n \xi_i(t) \exp\{\boldsymbol{\gamma}^T \mathbf{x}_i(t)\}},$$

and

$$\bar{y}(t, \boldsymbol{\gamma}) = \frac{\sum_{i=1}^n \xi_i(t) \exp\{\boldsymbol{\gamma}^T \mathbf{x}_i(t)\} y_i(t)}{\sum_{i=1}^n \xi_i(t) \exp\{\boldsymbol{\gamma}^T \mathbf{x}_i(t)\}},$$

where $\xi_i(t) = I(c_i \geq t)$. Replacing $dN_i(t)$ in (3.4) by its expectation (3.5), we obtain

$$\sum_{i=1}^{n} \int_0^{\infty} w(t)\{y_i(t) - \alpha(t) - \boldsymbol{\beta}^T \mathbf{x}_i(t)\}^2 \xi_i(t) \exp\{\boldsymbol{\gamma}^T \mathbf{x}_i(t)\}\, d\Lambda(t).$$

For each given $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, minimizing the above criterion function with respect to function $\alpha(t)$ is equivalent to minimizing it at each given time t. This results in estimating the baseline function by

$$\widehat{\alpha}(t; \boldsymbol{\beta}, \boldsymbol{\gamma}) = \bar{y}(t, \boldsymbol{\gamma}) - \boldsymbol{\beta}^T \bar{\mathbf{x}}(t, \boldsymbol{\gamma}). \tag{3.6}$$

Substituting $\alpha(t)$ with $\widehat{\alpha}(t; \boldsymbol{\beta}, \boldsymbol{\gamma})$ in (3.4) yields

$$\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \int_0^{+\infty} w(t)[\{y_i(t) - \bar{y}(t, \boldsymbol{\gamma})\} - \boldsymbol{\beta}^T \{\mathbf{x}_i(t) - \bar{\mathbf{x}}(t, \boldsymbol{\gamma})\}]^2\, dN_i(t). \tag{3.7}$$

The parameter $\boldsymbol{\gamma}$ can be consistently estimated by its moment estimator $\widehat{\boldsymbol{\gamma}}$, the solution to

$$\sum_{i=1}^{n} \int_0^{+\infty} \{\mathbf{x}_i(t) - \bar{\mathbf{x}}(t, \boldsymbol{\gamma})\}\, dN_i(t) = 0.$$

Substituting $\widehat{\boldsymbol{\gamma}}$ for $\boldsymbol{\gamma}$ in (3.7), an explicit form for $\widehat{\boldsymbol{\beta}}$ can be derived.

The weighted least-squares problem (3.7) requires that the processes $y_i(t)$ and $\mathbf{x}_i(t)$ are fully observable until the censoring time c_i. This is an unrealistic assumption. Thus, Lin and Ying (2001) replaced the processes by their corresponding values at the nearest time where their values are observed. While this helps in practical implementations of the procedure, it introduces biases due to the nearest neighborhood approximation. See Figure 1 for the within subject approximations. Furthermore, since for each subject the spaces among observation times $\{t_{ij}, j = 1, \cdots, J_i\}$ do not tend to zero even when the sample size n tends to infinity, the approximation biases cannot always be negligible in practice.

To improve efficiency and avoid nonnegligible biases due to approximations, Fan and Li (2004) proposed two estimators: a difference-based estimator and a profile least squares estimator. Dropping the subscript j, the observed data

$$\{(t_{ij}, \mathbf{x}(t_{ij})^T, \mathbf{y}(t_{ij})), j = 1, \cdots, J_i, i = 1, \cdots, n\},$$

can be expressed in the vector notation as

$$\{(t_i, \mathbf{x}_i^T, \mathbf{y}_i), i = 1, \cdots, n^*\}, \quad \text{with} \quad n^* = \sum_{i=1}^{n} J_i,$$

ordered according to the time $\{t_{ij}\}$. By the marginal model (3.3), it follows that

$$y_i = \alpha(t_i) + \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i, \quad \text{with} \quad E(\varepsilon_i|\mathbf{x}_i) = 0. \tag{3.8}$$

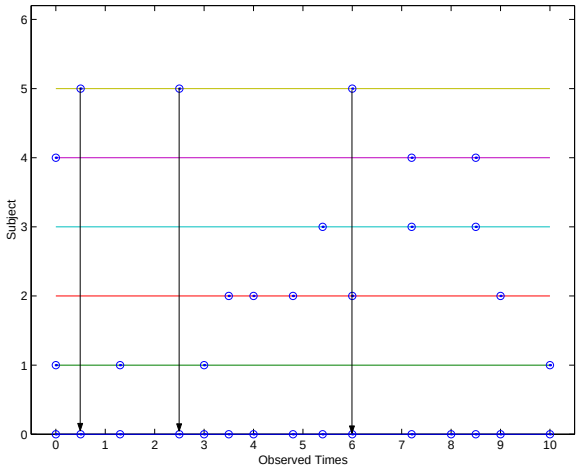

Fig. 1. Projection of observed time points

As illustrated in Figure 1, all observed times across all subjects may be dense and $t_{i+1} - t_i$ may be very small, although observed times for an individual subject may be very sparse. Observe that

$$y_{i+1} - y_i = \alpha(t_{i+1}) - \alpha(t_i) + \boldsymbol{\beta}^T(\mathbf{x}_{i+1} - \mathbf{x}_i) + e_i, \quad i = 1, \cdots, n^* - 1, \tag{3.9}$$

where stochastic error $e_i = \varepsilon_{i+1} - \varepsilon_i$. Under some mild conditions, the spacing between t_i and t_{i+1} is of order $O(1/n)$. Hence, the term $\alpha(t_{i+1}) - \alpha(t_i)$ in (3.9) is negligible. The least-squares approach can be employed to estimate the parameter $\boldsymbol{\beta}$. The method can be further improved by fitting the following linear model

$$y_{i+1} - y_i = \alpha_0 + \alpha_1(t_{i+1} - t_i) + \boldsymbol{\beta}^T(\mathbf{x}_{i+1} - \mathbf{x}_i) + e_i, \quad i = 1, \cdots, n^* - 1. \tag{3.10}$$

Fitting model (3.10) yields an estimate of $\boldsymbol{\beta}$. For simplicity, we will call this method the Difference Based Estimator (DBE). From simulation comparisons in Fan and Li (2004), the DBE outperforms Lin and Ying's approach. This is mainly due to the fact that within-subject nearest neighborhood approximations are much rougher than those in the pooled samples since

the former has much wider time gaps (See Fig. 1). We next present the profile least squares approach.

For a given $\boldsymbol{\beta}$, let $y^*(t) \equiv y(t) - \boldsymbol{\beta}^T \mathbf{x}(t)$. Then partially linear model (3.3) can be written as

$$y^*(t) = \alpha(t) + \varepsilon(t). \tag{3.11}$$

This is a nonparametric regression problem. Thus, one can use a nonparametric regression technique to estimate $\alpha(t)$. We will focus only on the local linear regression technique (Fan and Gijbels, 1996). For t in a neighborhood of t_0, it follows by the Taylor expansion that

$$\alpha(t) \approx \alpha(t_0) + \alpha'(t_0)(t - t_0) \equiv a_0 + a_1(t - t_0).$$

Let $K(\cdot)$ be a kernel function and h be a bandwidth. The local linear fit is to find $(\widehat{a}, \widehat{b})$ minimizing

$$\sum_{i=1}^{n} \sum_{j=1}^{J_i} \{y_i^*(t_{ij}) - a_0 - a_1(t_{ij} - t_0)\}^2 w(t_{ij}) K_h(t_{ij} - t_0). \tag{3.12}$$

Here the weight function, $w(t_{ij})$, serves a similar purpose to that in (3.4). The local linear estimate is simply $\widehat{\alpha}(t_0; \boldsymbol{\beta}) = \widehat{a}_0$.

We may derive succinct expression of the profile least squares estimator using matrix notation. Denote by $\mathbf{y}_i = (y_i(t_{i1}), \cdots, y_i(t_{iJ_i}))^T$, $\mathbf{y} = (\mathbf{y}_1^T, \cdots, \mathbf{y}_n^T)$, $\mathbf{X} = (\mathbf{X}_1^T, \cdots, \mathbf{X}_n^T)^T$ $\mathbf{X}_i = (\mathbf{x}_i(t_{i1}), \cdots, \mathbf{x}_i(t_{iJ_i}))^T$, $\boldsymbol{\alpha}_i = (\alpha(t_{i1}), \cdots, \alpha(t_{iJ_i}))^T$, and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \cdots, \boldsymbol{\alpha}_n^T)^T$. Then, model (3.8) can be written as

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.13}$$

where $\boldsymbol{\varepsilon}$ is the vector of stochastic errors. It is well known that the local linear fit is linear in $y_i^*(t_{ij})$ (Fan and Gijbels, 1996). Thus, the estimate of $\alpha(t)$ is linear in $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$. Hence, the estimate for the vector $\boldsymbol{\alpha}$ can be expressed as $\widehat{\boldsymbol{\alpha}} = \mathbf{S}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. The matrix $\mathbf{S}$ is usually called a smoothing matrix of the local linear smoother. It depends only on the observation times $\{t_{ij}, i = 1, \cdots, n, j = 1, \cdots, J_i\}$ and the amount of smoothing h. Substituting $\widehat{\boldsymbol{\alpha}}$ into (3.13), we obtain

$$(\mathbf{I} - \mathbf{S})\mathbf{y} = (\mathbf{I} - \mathbf{S})\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.14}$$

where $\mathbf{I}$ is the identity matrix of order $n^* = \sum_i n_i$. Applying weighted least-squares to the linear model (3.14), we obtain

$$\widehat{\boldsymbol{\beta}} = \{\mathbf{X}^T(\mathbf{I} - \mathbf{S})^T \mathbf{W}(\mathbf{I} - \mathbf{S})\mathbf{X}\}^{-1} \mathbf{X}^T(\mathbf{I} - \mathbf{S})^T \mathbf{W}(\mathbf{I} - \mathbf{S})\mathbf{y}, \tag{3.15}$$

where $\mathbf{W}$ is the weight matrix in the general least-squares, which can incorporate the within subject correlation. Working independence is also allowed, in which $\mathbf{W}$ is a diagonal matrix. The estimator in (3.15) is called the profile least-squares estimator. The profile least-squares estimator for the nonparametric component is simply $\alpha(\cdot; \widehat{\boldsymbol{\beta}})$. Fan and Li (2004) derived an estimate for the covariance matrix using a sandwich formula and discussed the issue of bandwidth selection. They suggested the following procedure for selecting a bandwidth.

Using (3.15) and noting that $\widehat{\boldsymbol{\alpha}} = \mathbf{S}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$, $\widehat{\boldsymbol{\alpha}}$ is linear in $\mathbf{y}$. Data driven methods, such as cross-validation (CV), or generalized cross-validation (GCV), can be used to select the bandwidth. However, this can computationally expensive. To avoid expensive computations and possibly unstable numerical implementations, our practical choice of bandwidth is as follows. Use the DBE to get an estimate $\widehat{\boldsymbol{\beta}}_{DBE}$. Substituting it into (3.11), we have a univariate nonparametric regression problem. Let $\widehat{h}$ be the bandwidth that is appropriate for this problem. This can be obtained either by a subjective choice via visualization, or by a data-driven procedure, such as substitution methods or cross-validation methods. Use this $\widehat{h}$ for the profile least-squares estimate. From nonparametric theory, this optimal choice of bandwidth is of order $h_n = bn^{-1/5}$. Fan and Li (2004) established the asymptotic normality of the profile least squares estimators. The performance of $\widehat{\boldsymbol{\beta}}$ is not very sensitive to the choice of h, namely, for a wide range of choice of bandwidth, the performance of $\widehat{\boldsymbol{\beta}}$ remains approximately the same. Liang, *et al.* (2004) consider estimation for the partially linear model with missing covariates.

3.2. *Variable selection*

Like parametric regression models, variable selection is important in the semiparametric model (3.3). The number of variables in (3.3) can easily be large when nonlinear terms and interactions between covariates are introduced to reduce possible modeling biases. It is common in practice to include only important variables in the model to enhance predictability and to give a parsimonious description of the relationship between the response and the covariates. Fan and Li (2004) proposed a class of variable selection procedure via the nonconvex penalized quadratic loss

$$\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\alpha}_i - \mathbf{X}_i\boldsymbol{\beta})^T W_i(\mathbf{y}_i - \boldsymbol{\alpha}_i - \mathbf{X}_i\boldsymbol{\beta}) + n\sum_{j=1}^{d}\lambda_j p_j(|\beta_j|).$$

where the $p_j(\cdot)$'s are penalty functions, and the λ_j's are tuning parameters, which control the model complexity and can be selected by some data-driven methods, such as cross validation or generalized cross validation.

After eliminating the nonparametric function $\alpha(\cdot)$ using the profiling technique [see (3.14)], we obtain the following penalized least squares:

$$\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}(\mathbf{I}-\mathbf{S})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+n\sum_{j=1}^{d}\lambda_j p_j(|\beta_j|). \tag{3.16}$$

By minimizing (3.16), with a special construction of the penalty function, some coefficients are estimated as zero, which deletes the corresponding variables, while others are not. Thus, the procedure selects variables and estimates coefficients simultaneously by minimizing (3.16). The resulting estimate is called a penalized least-squares estimate.

The penalty functions $p_j(\cdot)$ and the regularization parameters λ_j are not necessarily the same for all j. This allows us to incorporate prior information for the unknown coefficients by using different penalty functions or taking different values of λ_j. For instance, we may wish to keep important predictors in linear regression models and hence do not want to penalize their coefficients; so we take their $\lambda_j's$ to be zero. For ease of presentation, we denote $\lambda_j p_j(\cdot)$ by $p_{\lambda_j}(\cdot)$.

Many penalty functions, such as the family of L_q-penalty ($q \geq 0$), have been used for penalized least squares and penalized likelihood in various parametric models. Fan and Li (2001) provided various insights into how a penalty function should be chosen. and suggested the use of the smoothly clipped absolute deviation (SCAD) penalty. Its first derivative is defined by

$$p'_\lambda(\beta)=\lambda\left\{I(\beta\leq\lambda)+\frac{(a\lambda-\beta)_+}{(a-1)\lambda}I(\beta>\lambda)\right\},\ \text{for some } a>2 \text{ and } \beta>0,$$

and $p_\lambda(0)=0$. The SCAD penalty involves two unknown parameters, λ and a. Fan and Li (2001) suggested using $a=3.7$ from a Bayesian point of view.

Figure 2 depicts the plots of the SCAD, $L_{0.5}$ and L_1 penalty functions. As shown in Figure 2, the three penalty functions all are singular at the origin. This is a necessary condition for sparsity in variable selection: the resulting estimator automatically sets some small coefficients to be zero (Antoniadis and Fan, 2001). Furthermore, the SCAD and $L_{0.5}$ penalties are nonconvex over $(0,+\infty)$ in order to reduce estimation bias. We refer to penalized least squares with the nonconvex penalties over $(0,\infty)$ as *nonconvex penalized least squares* in order to distinguish from the L_2 penalty,

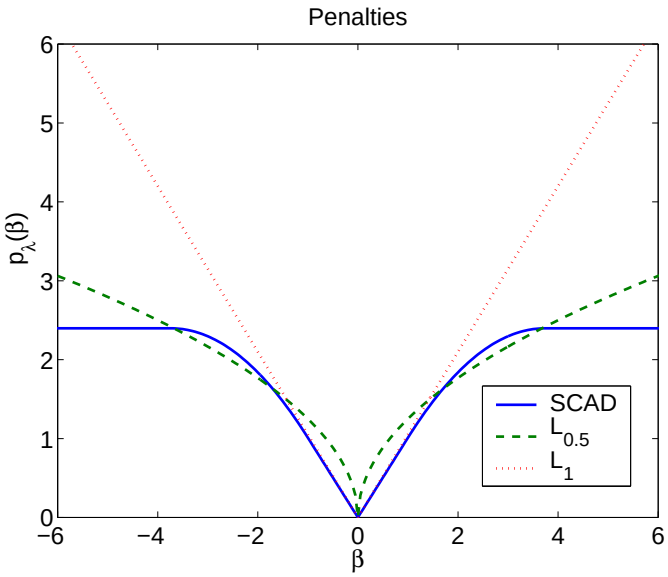

Fig. 2. Plot of Penalty Functions

which yields a ridge regression. The SCAD is an improvement over the L_0-penalty in two aspects: saving computational cost and resulting in a continuous solution to avoid unnecessary modeling variation. Furthermore, the SCAD improves bridge regression (using L_p-penalty) by reducing modeling variation in model prediction. Although similar in spirit to the L_1-penalty, the SCAD may improve the L_1-penalty by avoiding excessive estimation bias because the solution of the L_1-penalty always shrinks regression coefficients by a constant, for instance, the soft thresholding rule (Donoho and Johnstone, 1994 and Tibshirani, 1996). In contrast, the SCAD does not excessively over-penalize large coefficients.

Fan and Li (2001) suggested using local quadratic approximation for the nonconvex penalty functions, such as the SCAD penalty. With the aid of local quadratic approximation, Fan and Li (2004) proposed an iterative ridge regression algorithm for the penalized least squares (3.16). They further studied the sampling properties of the proposed variable selection procedures. They demonstrated that with a proper choice of regularization parameters and penalty functions, the proposed variable selection procedures perform as well as an oracle estimator.

3.3. *Extensions*

As an extension of partially linear models, Severini and Staniswalis (1994) considered generalized partially linear models. In the longitudinal data analysis, the following generalized partially linear models are usually considered

$$E\{y_i(t_{ij})|\mathbf{x}_i(t_{ij}\} = \mu\{\alpha(t_{ij}) + \mathbf{x}_i^T(t_{ij})\boldsymbol{\beta}\}. \tag{3.17}$$

Severini and Staniswalis (1994) proposed a profile quasi-likelihood estimation procedure for model (3.17). Lin and Carroll (2001a, b) developed estimation procedures for model (3.17) using profile-kernel estimation equation method. These authors have also studied the sampling properties of their proposed estimation procedures.

He, Zhu and Fung (2002) extended M-estimators for model (3.3). They approximate the baseline function $\alpha(t)$ by a regression spline. Thus, any M-estimation algorithm for the usual linear models can be used to obtain estimators of the model.

4. Varying-coefficient models

With the same notation in (3.3), the varying-coefficient model is defined as

$$y_i(t_{ij}) = \mathbf{x}_i^T(t_{ij})\boldsymbol{\beta}(t_{ij}) + \varepsilon_i(t_{ij}), \tag{4.1}$$

where we set the first elements of $\mathbf{x}_i(t_{ij})$ to be 1 to include an intercept function. Compared with model (3.3), the varying-coefficient model allows all regression coefficients to vary over time. Hence, it is also referred to as a time-varying coefficient model, which was introduced in Hastie and Tibshirani (1993). The main distinction between the functional data and longitudinal data setting is that the functional data in the current setting are observed at a much more frequently as if the whole functions were observed. Faraway (1997) considered model (4.1) with time-invariant covariate for functional data and proposed a smoothing splines procedure for the coefficients. Hoover, *et al.* (1998) and Wu, *et al.* (1998) are among the first to introduce the varying-coefficient models for analysis of longitudinal data.

Hoover, *et al.* (1998) proposed two estimation procedures for model (4.1). One is to globally approximate $\boldsymbol{\beta}(t)$ by splines and then estimate $\boldsymbol{\beta}(t)$ by smoothing spline. The other one is to locally approximate $\boldsymbol{\beta}(t)$ and to use local polynomial regression techniques to derive an estimator for $\boldsymbol{\beta}(t)$. Wu, *et al.* (1998) carefully studied the asymptotic properties of

the local polynomial regression estimator for $\boldsymbol{\beta}(t)$ and derived asymptotic confidence region for $\boldsymbol{\beta}(t)$. Faraway (1999) proposed a graphical method of exploring the mean structure using model (4.1). Wu and Chiang (2000) proposed a cross-validation criterion for selecting data-driven bandwidth and a bootstrap procedure for constructing confidence intervals.

Fan and Zhang (2000) proposed a two-step estimation procedure for model (4.1) for functional data. For simplicity of description, assume that (4.1) is observed at the same time and dense points $\{t_j, j = 1, \cdots, J\}$. This can be achieved by binning the functional data with respect to observed times t_{ij} if necessary. They suggested that in the first step, we estimate $\boldsymbol{\beta}(t_j)$ by linear regression using n data points collected at time t_j. Having estimated $\widehat{\boldsymbol{\beta}}(t)$ over $\{t_1, \cdots, t_J\}$, in the second step, they use the local polynomial regression to smooth $(\{(t_j, \widehat{\boldsymbol{\beta}}(t_j)), j = 1, \cdots, J\}$ componentwise. This procedure can be easily implemented and allows different coefficients to have different degrees of smoothness. Chiang, Rice and Wu (2001) applied the two-step estimation procedure to model (4.1) with time-invariant covariate and use smoothing splines in the second step.

Statistical inference on model (4.1) is still an active research area. Huang, Wu and Zhou (2002, 2004) proposed regression splines for varying coefficient models and studied the asymptotic properties of the resulting estimate. Eubank, *et al.* (2004) proposed smoothing spline estimators for inference in model (4.1) and developed Bayesian confidence intervals for the regression coefficient functions.

We can extend model (4.1) in the fashion of generalized linear models. This yields a generalized varying-coefficient model

$$E\{y_i(t_{ij})|\mathbf{x}_i(t_{ij})\} = \mu\{\mathbf{x}_i^T(t_{ij})\boldsymbol{\beta}(t)\}. \tag{4.2}$$

Cai, Fan and Li (2000) proposed efficient statistical inference procedures for the generalized varying coefficient model. Fan, Yao and Cai (2003) proposed adaptive varying coefficient model by allowing the link function to be unknown. Kauermann (2000) used model (4.2) to fit ordinal response longitudinal data. Qu and Li (2005) proposed a quadratic inference function approach (Qu, Lindsay and Li, 2000) to include within-subject correlation information into statistical inference on $\boldsymbol{\beta}(t)$.

Another extension of model (4.1) is the varying-coefficient mixed model

$$y_i(t_{ij}) = \mathbf{x}_i^T(t_{ij})\boldsymbol{\beta}(t_{ij}) + \mathbf{z}_i^T(t_{ij})\mathbf{b}_i(t_{ij}) + \varepsilon_i(t_{ij}), \tag{4.3}$$

where $\mathbf{b}_i(t)$ is random effect and both $\mathbf{x}_i$ and $\mathbf{z}_i$ are covariate vectors. The model considered in Wu and Zhang (2002) and Rice and Wu (2003) is a

specific case with $\mathbf{x}_i(t_{ij}) = 1$ and $\mathbf{z}_i(t_{ij}) = 1$. Wu and Liang (2004) proposed an estimation procedure for model (4.3) using local polynomial regression techniques. Li, Root and Shiffman (2005) applied model (4.3) for analysis of intensively correlated longitudinal data from a study of the subjective sensation of nicotine withdrawal. Zhang (2004) included a constant random effects in model (4.2) and obtained a generalized linear mixed effects model with time-varying coefficients. He further proposed an estimation procedure for his model using smoothing splines.

5. An illustration

We now illustrate the proposed procedures in Sections 2 and 3 via an analysis of a subset of data from the Multi-Center AIDS Cohort study. The data set contains the HIV status of 283 homosexual men who became infected with HIV during the follow-up period between 1984 and 1991. Details about the design, methods and medical implications of the study can be found in Kaslow *et al.*. (1987). During this study, all participants were scheduled to have their measurements taken during semi-annual visits. But, because many participants missed some of their scheduled visits, and the HIV infections happened randomly during the study, there are unequal numbers of repeated measurements and different measurement times per individual. Fan and Zhang (2000), and Huang, Wu and Zhou (2002) analyzed the same data set by using varying-coefficient models.

Take x_1 to be the smoking status: 1 for a smoker and 0 for a nonsmoker, $x_2(t)$ to be the standardized variable for age, and x_3 to be the standardized variable for PreCD4, the baseline CD4 percentage before HIV infection. It is of interest to examine whether there are any interaction effects and quadratic effects from these covariates. Based on the analysis of Huang, Wu and Zhou (2002), Fan and Li (2004) considered the following semiparametric model:

$$\begin{aligned} y(t) = \alpha(t) + \beta_1 x_1 + \beta_2 x_2(t) + \beta_3 x_3 + \beta_4 x_2^2(t) + \beta_5 x_3^2 \\ + \beta_6 x_1 x_2(t) + \beta_7 x_1 x_3 + \beta_8 x_2(t) x_3 + \varepsilon(t). \end{aligned} \tag{5.1}$$

The DBE estimate for $\boldsymbol{\beta}$ was computed to obtain the partial residuals for $\alpha(\cdot)$, and then the bandwidth $h = 0.5912$ was selected by the plug-in method proposed in Ruppert, Sheather and Wand (1995). After that, the profile least squares method with weight $w(t) \equiv 1$ was applied to this model. The resulting estimates and standard errors are depicted in Table 1. Figure 3 depicts the estimated baseline function $\alpha(t)$ along with its

Table 1. Estimated Coefficients for Model (5.1), adapted from Fan and Li (2004)

Variable	Profile LS $\widehat{\beta}(\text{se}(\widehat{\beta}))$	SCAD $\widehat{\beta}(\text{se}(\widehat{\beta}))$
Smoking	0.5333(1.0972)	0(0)
Age	-0.1010(0.9167)	0(0)
PreCD4	2.8252(0.8244)	3.1993(0.5699)
Age^2	0.1171(0.4558)	0(0)
PreCD4^2	-0.0333(0.3269))	0(0)
Smoking*Age	-1.7084(1.1192)	-1.0581(0.5221)
Smoking*PreCD4	1.3277(1.3125)	0(0)
Age*PreCD4	-0.1360(0.5413)	0(0)

95% pointwise confidence interval without taking account into the bias of the nonparametric fit. A decreasing trend can easily be seen, as the CD4 percentage depletes over time.

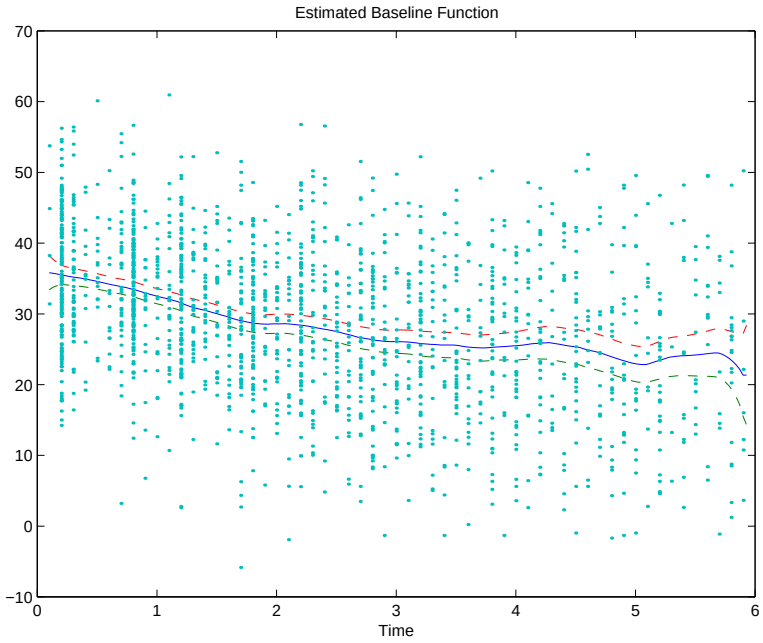

Fig. 3. Estimated Baseline Function. The solid line stands for the estimated baseline function, the dash lines are the estimated baseline function plus/minus twice standard errors. The dots are the residual on parametric part $r(t) = y(t) - \widehat{\boldsymbol{\beta}}^T \mathbf{x}(t)$. Taken from Fan and Li (2004).

Fan and Li (2004) further applied the penalized profile least squares approach to select significant variables. The generalized cross validation is used to select the tuning parameter and the selected $\lambda = 0.7213$ for the SCAD penalty. The results are also shown in Table 1. From Table 1, the result is in the line with that of Huang, Wu and Zhou (2002), but indicates possible interactions between Smoking status and Age.

6. Generalizations

A natural extension of varying-coefficient models and partially linear models is the semiparametric varying-coefficient model:

$$y_i(t_{ij}) = \mathbf{x}_i(t_{ij})^T \boldsymbol{\alpha}(t_{ij}) + \mathbf{z}_i(t_{ij})^T \boldsymbol{\beta} + \varepsilon_i(t_{ij}), \tag{6.1}$$

where both $\mathbf{x}_i(t)$ and $\mathbf{z}_i(t)$ are covariate vectors, $\boldsymbol{\alpha}(t)$ consists of p unknown smooth functions, $\boldsymbol{\beta}$ is a q-dimensional unknown parameter vector, and $E\{\varepsilon_i(t_{ij})|\mathbf{x}_i(t_{ij}), \mathbf{z}_i(t_{ij})\} = 0$. Martinussen and Scheike (1999) proposed an estimation procedure for model (6.1) using the notion of a counting process. Sun and Wu (2005) extended the estimation procedure of Lin and Ying for partially linear models to model (6.1). Fan, Huang and Li (2005) extended the profile least squares approach for model (6.1), and further proposed semiparametric modeling strategy for the covariance function of random error process $\varepsilon(t)$.

A generalization of model (6.1) is the semiparametric varying-coefficient mixed effects model:

$$y_i(t_{ij}) = \mathbf{x}_{i1}^T(t_{ij})\boldsymbol{\alpha}(t) + \mathbf{z}_{i1}^T(t_{ij})\boldsymbol{\beta} + \mathbf{x}_{i2}^T(t_{ij})\mathbf{a}_i^T(t_{ij}) + \mathbf{b}_i^T\mathbf{z}_{i2}(t_{ij}) + \varepsilon_i(t_{ij}), \tag{6.2}$$

where $\boldsymbol{\alpha}(t)$ consists of time-dependent fixed effects, $\boldsymbol{\beta}$ is time-independent fixed effects, $\mathbf{a}_i(t)$ is time-dependent random effects, and $\mathbf{b}_i$ is time-independent random effects. Fung, *et al.* (2002) considered partially linear mixed effects model which coincides with model (6.2) with $\mathbf{x}_{i1}(t_{ij}) = 1$ and $\mathbf{x}_{i2}(t_{ij}) = 1$.

In situations where the response variable $y(t)$ is discrete (e.g., binary, categorical, count), models (6.1) and (6.2) may not be appropriate. For such cases, one may consider the following further generalization

$$\begin{aligned} &E\{y_i(t_{ij})|\mathbf{x}_{i1}(t_{ij}), \mathbf{x}_{i2}(t_{ij}), \mathbf{z}_{i1}(t_{ij}), , \mathbf{z}_{i2}(t_{ij})\} \\ &= \mu\{\mathbf{x}_{i1}^T(t_{ij})\boldsymbol{\alpha}(t) + \mathbf{z}_{i1}^T(t_{ij})\boldsymbol{\beta} + \mathbf{x}_{i2}^T(t_{ij})\boldsymbol{\alpha}_i^T(t_{ij}) + \boldsymbol{\beta}_i^T\mathbf{z}_{i2}(t_{ij})\}, \end{aligned} \tag{6.3}$$

where $\mu(\cdot)$ is a known link function. It is of interest to develop statistical inference procedures for models (6.2) and (6.3) and their associated theory. Further research is needed in this area.

Other extensions of aforementioned models are semiparametric additive mixed models (Zhang, *et al.* 1998, Zhang and Lin, 2003), generalized additive models for longitudinal data (Berhane, and Tibshirani, 1998) and generalized additive mixed models (Lin and Zhang, 1999, Zhang and Davidian, 2004).

7. Estimation of covariance matrix

Estimation of covariance functions is an important issue in the analysis of longitudinal data. It features prominently in forecasting the trajectory of an individual response over time and is closely related with improving the efficiency of estimated regression coefficients. Challenges arise in estimating covariance functions due to the fact that longitudinal data are frequently collected at irregular and possibly subject-specific time points. Interest in issue has surged in the recent literature.

For parametric regression model in the analysis of longitudinal data, various efforts have been made to improve efficiency for estimating the regression coefficients. For instance, Liang and Zeger (1986) discussed how to incorporate correlation structure into the GEE framework, and showed that the resulting estimate is the most efficient when the working correlation equals the inverse of the actual correlation. Much attention has been paid to nonparametric regression analysis for longitudinal data in the recent literature. Wang (2003) proposed a marginal kernel GEE and showed that when the working correlation matrix equals the inverse of the true one, the resulting estimate is the most efficient. The marginal kernel GEE approach is extended to the generalized partial linear model in the seminal paper by Wang, Lin and Carroll (2005). All of these works indicate that the estimation of covariance function plays an important role in the analysis of longitudinal data.

Some estimation procedures for large covariance matrices have been proposed in the literature. Daniels and Kass (2001) proposed a Bayesian approach which places priors on a covariance matrix so as to shrink it toward some structures. Using an unconstrained and statistically meaningful reparametrization of the covariance matrix, Daniels and Pourahmadi (2002) further introduce more flexible priors with many parameters to control shrinkage. Pourahmadi (1999, 2000) proposed a flexible, data based parametric approach to formulating models for covariance matrices. Wu and Pourahmadi (2003) further proposed nonparametric estimation of large covariance matrices for balanced or nearly balanced longitudinal data, based

on Cholesky decomposition and using two-step estimation procedure (Fan and Zhang, 2000). The key idea of the series work by Pourahmadi and his coauthors is that the covariance matrix, denoted by Σ, of a zero mean random vector $\mathbf{z} = (z_1, \cdots, z_m)^T$ can be diagonalized by a lower triangular matrix constructed from the regression coefficients when z_t is regressed on its predecessors $z_1, \cdots, z_t$. Specifically, for $t = 2, \cdots, m$,

$$y_t = \sum_{j=1}^{t-1} \phi_{t,t-j} y_{t-j} + \varepsilon_t, \quad L\Sigma L^T = D, \tag{7.1}$$

where L and D are unique, L is a unit lower triangular matrix having ones on its diagonal and $-\phi_{ij}$ at its (i,j)th element for $j \leq t$, and D is diagonal with $\sigma_t^2 = \mathrm{Var}(\varepsilon_t)$ as its diagonal entries. The Cholesky decomposition (7.1) converts the constraint entries of Σ into two groups of unconstrainted regression and variance parameters given by $\{\phi_{tj}, t = 2, \cdots, m; j = 1, \cdots, t-1\}$ and $\{\log \sigma_1^2, \cdots, \log \sigma_m^2\}$, respectively. Let

$$\beta_{j,m}(t/m) = \phi_{t,t-j}, \ \sigma_m(t/m) = \sigma_t$$

Then the decomposition (7.1) yields

$$y_t = \sum_{j=1}^{t-1} \beta_{j,m}(t/m) y_{t-j} + \sigma_m(t/m)\varepsilon_t, \quad (t = 0, 1, ...), \tag{7.2}$$

which can be viewed as a time-varying coefficient model in which $\beta_{j,m}(\cdot)$ and $\sigma_m(\cdot)$ are assumed to be smooth function. This allows us to smooth along the subdiagonals of L. Wu and Pourahmadi (2003) directly applied two-step estimation for time-varying coefficient model (Fan and Zhang, 2000, also see Section 4 for a brief introduction) to (7.2). Using the Cholesky decomposition, Huang, Liu and Pourahmadi (2005) further introduced a penalized likelihood method for estimating a large covariance matrix. Yao, Müller and Wang (2005a, b) proposed other new estimation procedures for the covariance function of functional data.

Diggle and Verbyla (1998) proposed using local linear regression techniques to estimate covariance structure for longitudinal data. Their approach is to estimate variance function and variogram by smoothing the squared residual and the variogram cloud of the squares residuals. Then using the relationship between variogram and covariance function, one may derive an estimation for covariance function. However, the resulting estimate for covariance function may not be positive definite. Fan, Huang and Li (2005) gave a method for parsimonious modeling of the covariance function of random error process $\varepsilon(t)$ for the analysis of longitudinal data when

they are collected at irregular and possibly subject-specific time points. They approach this by assuming that $\mathrm{Var}\{\varepsilon(t)|\mathbf{x}(t), \mathbf{z}(t)\} = \sigma^2(t)$, which is a nonparametric smoothing function, but the correlation function between $\varepsilon(s)$ and $\varepsilon(t)$ has a parametric form $\mathrm{corr}\{\varepsilon(s), \varepsilon(t)\} = \rho(t, s, \boldsymbol{\theta})$, where $\rho(s, t, \boldsymbol{\theta})$ is a positive definite function of s and t, and $\boldsymbol{\theta}$ is an unknown parameter vector. Specification of the correlation function may be motivated from the random error structure in hierarchical linear models and mixed effects models, or by specifying a working correlation function as in GEE (Liang and Zeger, 1986). For instance, an AR(1)-type correlation structure yields $\rho(s, t, \theta) = \exp(-\theta|s - t|)$ with $\theta > 0$, and a compound symmetric correlation structure results in $\rho(s, t, \theta) = \theta$ with $0 < \theta < 1$, which is the same as the correlation structure of random effects model in the presence of only random intercept. More complicated correlation structures can be introduced by using ARMA models or adopting a hierarchical linear model including various levels of random effects.

In Fan, Huang and Li (2005), the covariance function is fitted by a semiparametric model. The semiparametric model allows random error process $\varepsilon(t)$ to be nonstationary as its variance function $\sigma^2(t)$ may be time-dependent. The semiparametric model to the covariance function guarantees the positive definite property of the resulting estimate and retains the flexibility of nonparametric modeling and the parsimony of parametric modeling. Furthermore, this semiparametric model allows one to incorporate easily prior information about the correlation structure. It can be used to improve the efficiency of regression coefficient $\boldsymbol{\beta}$ even if the correlation matrix is misspecified. For example, let $\rho_0(s, t)$ be a working correlation function (e.g. working independence) and $\rho(s, t, \boldsymbol{\theta})$ be a family of correlation functions that contains ρ_0, the semiparametric model allows one to choose an appropriate $\boldsymbol{\theta}$ to improve the efficiency of the estimator of $\boldsymbol{\beta}$. Obviously, to improve the efficiency, the family of correlation functions $\{\rho(s, t, \boldsymbol{\theta})\}$ does not need to contain the true correlation structure even if the correlation matrix is misspecified. Fan, Huang and Li (2005) suggested estimating the variance function by using a kernel smoothing over the squares of residuals. They further proposed two estimation procedures for $\boldsymbol{\theta}$. One is motivated by maximizing the likelihood of the data, and the other is motivated by minimizing the volume of confidence ellipsoid of regression coefficients.

Acknowledgements

Fan's research was partially supported by DMS-0354223 and NIH grant R01 HL69720. Li's research was supported by a National Institute on Drug Abuse (NIDA) grant P50 DA10075 and a NSF grant DMS-0348869.

References

1. Altman, N.S. (1990). Kernel smoothing of data with correlated errors. *Journal of American Statistical Association*, **85**, 749-759.
2. Antoniadis, A. and Fan, J. (2001). Regularization of wavelets approximations (with discussions). *Journal of American Statistical Association*, **96**, 939-967.
3. Berhane, K. and Tibshirani, R. J. (1998). Generalized additive models for longitudinal data. *Canadian Journal of Statistics*, **26**, 517-535.
4. Cai, Z, Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of American Statistical Association*, **95**, 888-902.
5. Carroll, R. J., Hall, P., Apanasovivh, T.V. and Lin, X. (2004). Histospline method in nonparametric regression models with application to clustered longitudinal data. *Statistica Sinica*, **14**, 649-674.
6. Chiang, C.-T., Rice, J.A. and Wu, C.O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of American Statistical Association*, **96**, 605-619.
7. Daniels, M. J. and Kass, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics*, **57**, 1173-1184.
8. Daniels, M. J. and Pourahmadi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, **89**, 553-566.
9. Diggle, P.J., Heagerty, P., Liang, K.Y., Zeger, S.L. (2002). *Analysis of Longitudinal Data*, 2nd Edition, Oxford, U. K. Oxford University Press.
10. Diggle, P.J. and Verbyla, A.P. (1998). Nonparametric estimation of covariance structure of longitudinal data. *Biometrics*, **54**, 401-415.
11. Engle, R.F., Granger,C.W.J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of American Statistical Association*, **81**, 310-320.
12. Eubank, R.L., Huang, C.F., Maldonado, Y.M., Wang, N., Wang, S. and Buchanan, R.J. (2004). Smoothing spline estimation in varying-coefficient models *Journal of Royal Statistical Society, Series B*, **66**, 653-667.
13. Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman and Hall, London.
14. Fan, J., Heckman, N.E. and Wand, M.P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of American Statistical Association*, **90**, 141–150.
15. Fan, J., Huang, T. and Li, R. (2005). Analysis of longitudinal data with semiparametric estimation of covariance function. Manuscript.

16. Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, **96**, 1348-1360.
17. Fan, J and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis, *Journal of American Statistical Association*, **99**, 710-72.
18. Fan, J. and Zhang, J. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of Royal Statistical Society, Series B*, **62**, 303-322.
19. Faraway, J. J. (1997). Regression analysis for a functional response. *Technometrics*, **39**, 254-261.
20. Faraway, J.J. (1999). A graphical method of exploring the mean structure in longitudinal data analysis. *Journal of Computational and Graphical Statistics*, **8**, 60-68.
21. Fung, W.K., Zhu, Z.Y., Wei, B.C. and He, X.M. (2002). Influence diagnostics and outlier tests for semiparametric mixed models. *Journal of Royal Statistical Society, Series B*, **64** 565-579.
22. Härdle, W. Liang, H. and Gao, J. (1999). *Partially Linear Models*, Springer-Verlag, New York.
23. Hart, J.D. (1991). Kernel regression estiamtion with time-series errors. *Journal of Royal Statistical Society, Series B*, **53**, 173-187.
24. Hart, J.D. and Wehrly, T.E. (1986). Kernel regression estimation using repeated measurements data. *Journal of American Statistical Association*, **81**, 1080-1088.
25. Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models (with discussion). *Journal of Royal Statistical Society, Series B*, **55**, 757-796.
26. He, X.M., Zhu, Z.Y., and Fung, W.K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, **89**, 579-590.
27. Heckman, N. (1986). Spline smoothing in partly linear models, *Journal of Royal Statistical Society, Series B*, **48**, 244-248.
28. Hoover, D.R., Rice, J.A., Wu, C.O. and Yang, L.P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809–822.
29. Huang, J.Z., Liu, N. and Pourahmadi, M. (2005). Covariance selection and estimation via penalized normal likelihood. *Biometrika*. To appear.
30. Huang, J.Z., Wu, C. O. and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, **89**, 111-128.
31. Huang, J. Z, Wu, C.O. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, **14**, 763-788.
32. Kaslow, R.A., Ostrow, D.G., Detels, R. Phair, J.P., Polk, B.F. and Rinaldo, C.R. (1987). The Multicenter AIDS Cohort Study: rationale, organization and selected characteristics of the participants. *Am. J. Epidem.*, **126**, 310-318.

33. Kauermann G (2000). Modeling longitudinal data with ordinal response by varying coefficients, *Biometrics*, **56**, 692-698.
34. Li, R., Root, T. and Shiffman, S. (2005). A local linear estimation procedure for functional multilevel modeling. In *Models for Intensively Longitudinal Data*, (T. Walls and J. Schafer eds), 63-83. Oxford University Press.
35. Liang, H., Wang, S.J., Robins, J.M. and Carroll, R.J. (2004). Estimation in partially linear models with missing covariates *Journal of American Statistical Association*, **99**, 357-367.
36. Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22
37. Lin, D.Y. and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data (with discussion). *Journal of American Statistical Association*, **96**, 103-126.
38. Lin, X. and Carroll, R.J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of American Statistical Association*, **95**, 520-534.
39. Lin, X. and Carroll, R.J. (2001a). Semiparametric regression for clustered data. *Biometrika*, **88**, 1179-1185.
40. Lin, X, and Carroll, R.J. (2001b). Semiparametric regression for clustered data using generalized estimating equations. *Journal of American Statistical Association*, **96**, 1045-1056.
41. Lin, X., Wang, N.Y., Welsh, A.H, and Carroll, R.J. (2004). Equivalent kernels of smoothing splines in nonparametric regression for clustered/longitudinal data. *Biometrika*, **91**, 177-193.
42. Lin, X. and Zhang, D. (1999) Inference in generalized additive mixed models by using smoothing splines. *Journal of Royal Statistical Society, Series B*, **61**, 381-400.
43. Martinussen, T. and Scheike, T.H. (1999). A semiparametric additive regression model for longitudinal data. *Biometrika*, **86**, 691-702.
44. Moyeed, R.A. and Diggle, P.J. (1994). Rates of convergence in semiparametric modeling of longitudinal data. *Austr. Jour. Statist.*, **36**, 75-93.
45. Müller, H.G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*. Springer-Verlag, New York.
46. Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterization. *Biometrika*, **86**, 677-690.
47. Pourahmadi, M. (2000). Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix. *Biometrika*, **87**, 425-435.
48. Peterson, D.R., Zhao, H. and Eapen, S. (2003). Using local correlation in kernel-based smoothers for dependent data. *Biometrics*, **59**, 984-991.
49. Qu, A. and Li, R. (2005). Quadratic inference functions for varying coefficient models with longitudinal data, *Biometrics*. To appear.
50. Qu, A., Lindsay, B. G., and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika*, **87**, 823-836.
51. Rice, J.A. and Wu, C.O. (2003). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253-259.

52. Ross, S.M. (1997). *Simulation*. Second Edition. Academic Press, Inc., San Diego, CA.
53. Ruckstuhl, A.,Welsh, A. and Carroll, R.J. (2000). Nonparametric function estimation of the relationship between two repeatedly measured variables. *Statistica Sinica*, **10**, 51-71.
54. Ruppert, D., Sheather, S.J. and Wand, M.P. (1995). An effective bandwidth selector for local least squares regression. *Journal of American Statistical Association*, **90**, 1257-1270.
55. Severini, T.A. and Staniswalis, J.G. (1994). Quasi-likelihood estimation in semiparametric models. *Journal of American Statistical Association*, **89**, 501–511.
56. Silverman, B. (1984). Spline smoothing: the equivalent variable kernel method. *Annals of Statistics*, **12**, 501-511.
57. Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of Royal Statistical Society, Series B*, **50**, 413–436.
58. Sun, Y. and Wu, H. (2005). Semiparametric time-varying coefficients regression model for longitudinal data. *Scandinavian Journal of Statistics*, **32**, 21 – 47.
59. Wahba, G. (1984). Partial spline models for semiparametric estimation of functions of several variables, in *Statist. Analysis of Time Ser.*, Proceedings of the Japan U.S. Joint Seminar, Tokyo, 319–329.
60. Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika*, **90**, 43-52.
61. Wang, N., Carroll, R.J. and Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of American Statistical Association*, **100**, 147-157.
62. Welsh, A.H., Lin, X. and Carroll, R.J. (2002). Marginal longitudinal nonparametric regression: Locality and efficiency of spline and kernel methods. *Journal of American Statistical Association*, **97**, 482-493.
63. Wu, C.O. and Chiang, C.T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica*, **10**, 433-456.
64. Wu, C.O., Chiang, C.T. and Hoover, D.R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of American Statistical Association*, **93**, 1388-1402.
65. Wu, H. and Liang, H. (2004). Random Varying-Coefficient Models with Smoothing Covariates, Applications to an AIDS Clinical Study. *Scan. J. Statist.*, **31**, 3-19.
66. Wu, F. and Zhang, J. (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of American Statistical Association*, **97** , 883-897.
67. Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, **90**, 831–844.
68. Xiao, Z., Linton, O.B., Carroll, R.J, and Mammen, E. (2003). More efficient local polynomial estimation in nonparametric regression with autocorrelated errors. *Journal of American Statistical Association*, **98**, 980-992.
69. Yao, F., Müller, H.G. and Wang, J.-L. (2005a). Functional data analysis for

sparse longitudinal data. *Journal of the American Statistical Association*, **100**, 577-590.

70. Yao, F., Müller, H.G., Wang, J.-L. (2005b). Functional Regression Analysis for Longitudinal Data. *The Annals of Statistics*, in press.
71. Zeger, S.L. and Diggle, P.J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, **50**, 689-699.
72. Zhang, D. (2004). Generalized linear mixed models with varying coefficients for longitudinal data. *Biometrics*, **60**, 8-15.
73. Zhang, D. and Davidian, M. (2004). Likelihood and conditional likelihood inference for generalized additive mixed models for clustered data. *Journal of Multivariate Analysis*, **91**, 90-106.
74. Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, **4**, 57-74.
75. Zhang, D., Lin, X., Raz, J., and Sowers, M.F. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of American Statistical Association*, **93**, 710-719.
76. Zhang, D., Lin, X. and Sowers, M.F. (2000). Semiparametric regression for periodic longitudinal hormone data from multiple menstrual cycles. *Biometrics*, **56**, 31-39.