Statistica Sinica **19** (2009), 1137-1153

A DYNAMIC QUANTILE REGRESSION TRANSFORMATION MODEL FOR LONGITUDINAL DATA

Yunming Mu and Ying Wei

Portland State University and Columbia University

Abstract: This paper describes a flexible nonparametric quantile regression model for longitudinal data. The basic elements of the model consist of a time-dependent power transformation on the longitudinal dependent variable and a varying-coefficient model for conditional quantiles. A two-step estimation procedure is proposed to fit the model, and its consistency is established. Tuning parameters are chosen with generalized cross validation in conjunction with a Schwarz-type information criterion. The proposed method is illustrated by data on the time evolution of CD4 cell counts in HIV-1 infected patients under three different treatments. The quantile regression approach for longitudinal data enables construction of a pointwise prediction band for CD4 cell counts trajectories without requiring parametric distributional assumptions.

Key words and phrases: Longitudinal data, power transformation, quantile regression, varying-coefficient models.

1. Introduction

The varying-coefficient model proposed by Hastie and Tibshirani (1993) extends the framework of classical generalized linear models. These models are particularly appealing in longitudinal studies, where they allow one to explore the extent to which covariates affect responses changing over time. Such models have been extensively studied in the literature, see Hoover, Rice, Wu and Yang (1998), Fan and Zhang (2000), Wu and Chiang (2000), Wu, Yu and Chiang (2000) and Chiang, Rice and Wu (2001), among others. The varying-coefficient model assumes that the response is linearly related to its covariates at any time point. Due to the dynamic nature of many applications, this linear association may not hold at all time points. To relax the strict global linearity assumption, we extend the class of varying-coefficient models by incorporating a time-dependent transformation on the longitudinal response to recover possible nonlinear patterns. On the other hand, virtually all the aforementioned work has focused on the problem of conditional mean and variance estimation, leaving behind other aspects of conditional distributions such as quantiles. When data exhibit skewness

or heavy-tails, conditional quantile model inference can uncover important features that would be overlooked by a mean-based analysis. In fact, quantile-based inference has been proven to be an effective tool in analyzing many longitudinal data sets. We refer to Wei, Pere, Koenker and He (2005) and Wei and He (2006) for quantile regression methods in constructing reference charts in medicine, and to Lipsitz, Fitzmaurice, Molenberghs and Zhao (1997) for analysis of CD4 cell counts data. Here we consider a model for longitudinal data under the quantile regression framework. The approach is through the class of marginal models.

Let $Y(t)$ and $X(t)$ be the positive real-valued outcome of interest and the R^p-valued column covariate vector, respectively, when observed at time t. We consider an experiment with m subjects and n_i observations over time for the ith subject $(i = 1, \ldots, m)$ for a total of $n = \sum_{i=1}^{m} n_i$ observations. The jth observation of $(t, X(t), Y(t))$ for the ith subject is denoted by (t_{ij}, x_{ij}, y_{ij}) for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, where x_{ij} is given by the column vector $x_{ij} = (x_{ij1}, \ldots, x_{ijp})$. Let $Q_Y(\tau|t, x)$ denote the τth quantile of the conditional distribution of Y given $X = x$ at time t. For any function $\lambda(t)$, let $\Lambda(y, t; \lambda) = (y^{\lambda(t)} - 1)/\lambda(t)$. We assume that the full dataset (t_{ij}, x_{ij}, y_{ij}), $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, is observed and can be modelled as

$$\Lambda(Y_{ij}, t_{ij}; \lambda_\tau) = \sum_{k=1}^{p} X_{ijk} \beta_{\tau,k}(t_{ij}) + e_{ij}, \tag{1.1}$$

where $\lambda_\tau(t)$ and $\beta_{\tau,k}(t)$ are arbitrary smooth functions of t, and e_{ij} satisfies the quantile constraint $Q_{e_{ij}}(\tau|t_{ij}, x_{ij}) = 0$. Model (1.1) implies that at any time t, the τth conditional quantile of the transformed response variable is a linear function of its covariates and the coefficients vary over time in an arbitrary form. By letting $x_{ij1} = 1$, the model allows a time-varying intercept term. We assume without loss of generality that the t_{ij} are all scaled into the interval $[0, 1]$. We further assume that the observations, and therefore the e_{ij}, from different subjects are independent. The form of the error distribution and of the within-subject correlations are not specified. Recently, a simple varying-coefficient model for conditional quantiles where no transformation is used on the response variable has been considered by Honda (2004) and Kim (2007) for independent cross-sectional data, and by Cai and Xu (2008) for time series data.

The proposed model is particularly useful for predicting the longitudinal response trajectory under minimal assumptions, since the model has flexibility in two ways: (1) it generalizes a simple vary-coefficient model by allowing nonlinearity at any time t; (2) it does not assume any parametric, particularly Gaussian, error distributions. A HIV dataset is used to illustrate potential applications of the method in Section 3.

One parametric approach for estimating quantile functions using transformations on $Y(t)$ is known as the LMS method, originally proposed by Cole and Green (1992). The LMS method applies time-varying Box-Cox power transformations to transform $Y(t)$ to have the standard normal distribution, i.e., $Z(t) = ([Y(t)/\mu(t) - 1]^{1/\lambda(t)})/(\sigma(t)\lambda(t)) \sim N(0,1)$, where the parameter functions $\lambda(t)$, $\mu(t)$, and $\sigma(t)$ are estimated via maximizing a penalized likelihood function. The normality assumption after power transformations may lead to bias under certain circumstances, as illustrated in Wei and He (2006). In addition, incorporating additional covariates other than time is computationally difficult for the LMS method; extensions of the LMS method that allow covariates other than time have been considered in the literature, see Yee and Wild (1996) and Yee (2004). We compare our method with the method in Yee (2004) in several respects, using a HIV dataset in Section 3.

This paper is organized as follows. Section 2 describes a two-step estimation procedure of the model at (1.1), discusses the choice of smoothing parameters, and presents the main result of this paper that establishes the consistency of the proposed estimation procedure. In Section 3, we illustrate the usefulness of the proposed methods using a HIV dataset. A Monte Carlo simulation study is presented in Section 4. Section 5 concludes.

2. Estimation

A Two-Step Estimation Procedure. In this section, we introduce a two-step estimation procedure for estimating parameters in the model at (1.1). The proposed method is straightforward to implement and admits different degrees of smoothness of $\lambda_\tau(t)$ and $\beta_{\tau,k}(t)$'s. Consistency of the two-step procedure is established at the end of this section.

At the first step, we obtain raw estimates of $\lambda_\tau(t)$ on a subdivision of the range of t. At the second step, a smoother is applied to the raw estimates to produce final estimates of $\lambda_\tau(t)$, and the $\beta_{\tau,k}(t)$'s are estimated nonparametrically given the estimated $\lambda_\tau(t)$. To explicitly define the estimator, let k_n be a positive integer and partition the unit interval into k_n subintervals of the form $I_l = [(l-1)/k_n,\ l/k_n)$, $l = 1, \ldots, k_n - 1$, and $I_{k_n} = [(k_n - 1)/k_n, 1]$. At each subinterval I_l, $l = 1, \ldots, k_n$, we approximate the transformation function by a constant and apply the estimation method proposed by Mu and He (2007) to obtain a raw estimate of $\lambda_\tau(t)$, which we denote by $\tilde{\lambda}_{l,n}$. Then a smoother is applied to the raw estimates to produce the final estimates, denoted by $\hat{\lambda}_n(t)$. Finally we estimate the $\beta_{\tau,k}(t)$'s using regression splines assuming the estimated transformation function is given. We next introduce notation and detail the estimation procedure.

For a fixed $\lambda \in R$, define a cusum process of residuals on each subinterval I_l, $l = 1, \ldots, k_n$, as

$$R_{nl}(x, \lambda, \beta) = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathrm{I}\{t_{ij} \in I_l\} \mathrm{I}\{x_{ij} \leq x\} \left[\tau - \mathrm{I}\left\{ \frac{y_{ij}^{\lambda} - 1}{\lambda} - x_{ij}^T \beta \leq 0 \right\} \right], \quad (2.1)$$

where $\mathrm{I}\{\cdot\}$ is the indicator function and $\mathrm{I}\{x_{ij} \leq x\} = \mathrm{I}\{x_{ij1} \leq x_1, \ldots, x_{ijp} \leq x_p\}$. Let $\tilde{\beta}_{l,n}(\lambda)$ be the solution to the optimization problem

$$\min_{b \in R^p} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathrm{I}\{t_{ij} \in I_l\} \rho_\tau \left(\frac{y_{ij}^{\lambda} - 1}{\lambda} - x_{ij}^T b \right). \quad (2.2)$$

Associated with each subinterval I_l, we set

$$V_{nl}(\lambda) = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathrm{I}\{t_{ij} \in I_l\} \cdot \left[R_{nl}(x_{ij}, \lambda, \tilde{\beta}_{l,n}(\lambda)) \right]^2. \quad (2.3)$$

Then on each subinterval I_l, $l = 1, \ldots, k_n$, $\lambda_\tau(t)$ is estimated by a constant: $\tilde{\lambda}_{l,n} = \mathrm{argmin}_{\lambda \in \Omega_\tau} V_{nl}(\lambda)$. Notice that $\tilde{\beta}_{l,n}(\lambda)$, $R_{nl}(x, \lambda, \beta)$, and $V_{nl}(\lambda)$ all depend on τ, but we that drop the subscript τ for ease of presentation. A raw estimator for the unknown function $\lambda_\tau(t)$ at the first step can be expressed as

$$\tilde{\lambda}_n(t) = \sum_{l=1}^{k_n} \tilde{\lambda}_{l,n} \mathrm{I}\{t \in I_l\}. \quad (2.4)$$

At the second step, we refine the piecewise constant estimates via smoothing splines. Let $t_{(l)}$ denote the middle value of the interval $[t_{l-1}, t_l)$, and let $W_2[0,1]$ be the class of all functions that are continuously differentiable on the interval $[0,1]$ and have a second derivative that is square integrable on $[0,1]$. A smoothing spline estimator of $\lambda_\tau(t)$ is

$$\hat{\lambda}_n(t) = \mathrm{argmin}_{\lambda(\cdot) \in W_2[0,1]} \frac{1}{k_n} \sum_{l=1}^{k_n} (\tilde{\lambda}_{l,n} - \lambda(t_{(l)}))^2 + \gamma \int_0^1 [\lambda''(t)]^2 dt, \quad (2.5)$$

where γ controls the amount of smoothing. Notice that if the transformation function $\lambda_\tau(t)$ were known, model (1.1) reduces to a simple varying-coefficient quantile regression model. Thus, after an estimator for $\lambda_\tau(t)$ has been obtained, we can estimate $\beta_{\tau,k}(t)$ using one of the existing techniques for a simple varying-coefficient quantile regression model, for example the B-spline estimator used in Kim (2007) or the local polynomials approach adopted by Honda (2004). Here we adopt the B-spline approach to illustrate the idea. Let $\{b_l : l = 1, 2, \ldots\}$

denote a basis for smooth functions on [0, 1], and θ a m_n-dimensional vector. Define $B(t) = [1, b_1(t), \ldots, b_{m_n}(t)]^T$. The regression spline estimators of $\beta_{\tau,k}(t)$ are $\hat{\beta}_{n,k}(t) = B(t)^T\hat{\theta}_{n,k}$, $k = 1, \ldots, p$, where $\hat{\theta}_{n,k}$ solves the minimization problem

$$\text{argmin}_{\theta_k \in R^{m_n}, k=1\ldots,p} \sum_{i=1}^{m}\sum_{j=1}^{n_i} \rho_\tau(\Lambda(y_{ij}, t_{ij}; \hat{\lambda}_n) - x_{ij1}B(t_{ij})^T\theta_1 \cdots - x_{ijp}B(t_{ij})^T\theta_p). \tag{2.6}$$

Having estimated $\lambda_\tau(t)$ and $\beta_{\tau,k}(t)$, $k = 1, \ldots, p$, an estimate of the conditional τth quantile of Y_{ij} given $X_{ij} = x_{ij}$ at time t_{ij} is obtained as

$$\hat{Q}_{Y_{ij}}(\tau|t_{ij}, x_{ij}) = S(\hat{\lambda}_n(t_{ij}), \hat{q}(t_{ij}, x_{ij})), \tag{2.7}$$

where $S(\lambda, u) = (\lambda \cdot u + 1)^{1/\lambda}$ $(= \exp(u)$ if $\lambda = 0)$ denotes the inverse power transformation function, and $\hat{q}(t_{ij}, x_{ij}) = \hat{\beta}_{n,1}(t_{ij})x_{ij1} + \cdots + \hat{\beta}_{n,p}(t_{ij})x_{ijp}$.

Choice of Tuning Parameters. The empirical performance of the quantile estimator defined in (3.4) depends on several things: k_n, the number of subintervals on which we obtain raw estimates of the transformation function; γ, the smoothing parameter in (2.5); the order and knots for B-splines used in (2.6). In practice, under-smoothing or over-smoothing is mainly caused by inappropriate choices of γ in (2.5), but is rarely influenced by the number of subdivisions, k_n. Our simulation study suggests that the performance of the proposed method is quite stable over a wide range of values for k_n. However the method is designed for relatively large sample sizes, and we recommend choosing k_n between 15 and 50, and allowing at least 50 observations in each subinterval I_l, $l = 1, \ldots, k_n$.

The method is somewhat sensitive to the degree of smoothing of the transformation function. Common selection methods of γ in the smoothing spline literature can be used, say GCV, CV, etc. Because the raw estimates may be correlated, the GML method specially designed for correlated data might be used, see Wang (1998). A small simulation study shows that the GML method without correlations works slightly better than the GCV method in several settings. However the GML method tends to be computationally unstable when a correlation structure is imposed, say AR(1). In this paper we use the GCV method due to its popularity in routine applications, and its satisfactory performance in our simulation studies. However no automatic selection criterion is perfect in the real world, and sometimes a subjective choice based on observation of the data works well. In real data analysis, we recommend that the smoothed transformation curve be examined to check for artificial patterns, and that the smoothing parameter be manually adjusted if there is evidence of over-smoothing or under-smoothing.

When it comes to the order of B-splines used in constructing the basis functions for estimating $\beta_{\tau,k}(t)$'s, we suggest using lower order splines such as linear and quadratic splines. Since the effect of the splines on the model is multiplicative, higher order splines would induce complicated interactions and collinearity among the variables in the model. In this paper we use quadratic splines but linear splines can be used if we think that the coefficient functions are less smooth. Critical to the quality of a B-spline approximation is the selection of knots. We start with a set of knots equally spaced in percentile ranks by taking $s_l = t_{[ln/m_n]}$, the l/m_nth quantile of the distinct variables of t_i for $l = 1, \ldots, m_n$. We obtain m_n as the minimizer to the Schwarz-type information criterion

$$\mathrm{IC}(m) = \log\left(\sum_i \sum_j \rho_\tau(r_{ij})\right) + \frac{1}{2} n^{-1} \log(n) \cdot p_n, \tag{2.8}$$

where $r_{ij} = \Lambda(y_{ij}, t_{ij}; \hat{\lambda}_n) - x_{ij1} B(t_{ij})^T \hat{\theta}_{n,1} \cdots - x_{ijp} B(t_{ij})^T \hat{\theta}_{n,p}$, and $p_n = m + ord + 1$, with *ord* denoting the order of B-spline basis functions. Here we use a set of knots equally spaced in percentile ranks but a candidate set of knots provided by the user can also be used. After the number m_n has been determined, a stepwise knot selection method could be used to further select knots. A Wald statistic or Rao score statistic can be used to add or delete one knot at a time. Such a procedure is computationally intensive and is not our consideration here. Stepwise knot selection methods have been used by a number of authors, including Friedman and Silverman (1989), He and Shi (1996) and others.

Consistency. We list and discuss a set of assumptions that guarantees consistency of the estimated quantile estimator in (3.4). The main result of this paper is presented in Theorem 1; detailed proofs are provided in an online supplement available at the following URL `http://www.stat.sinica.edu.tw/statistica`.

Theorem 1. *If Assumptions* 1−10 *hold,* $h_n \to 0$ *and* $nh_n \to \infty$*, then*

$$\sup_{t \in [0,1]} |\hat{\lambda}_n(t) - \lambda_\tau(t)| = o_p(1), \tag{2.9}$$

$$\frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\hat{\beta}_{n,k}(t_{ij}) - \beta_{\tau,k}(t_{ij}))^2 = o_p(1), \quad k = 1, \ldots, p. \tag{2.10}$$

Assume that the measurement times T_{ij} are i.i.d with marginal distribution D_T and marginal density d_T. We first introduce smoothness conditions on the unknown transformation function and the coefficient functions in model (1.1).

Assumption 1. $\lambda_\tau(t)$ is twice continuously differentiable on $[0, 1]$. For each $k = 1, \ldots, p$, $\beta_{\tau,k}(t)$ is r_k times continuously differentiable on [0, 1] for some $r_k \geq 1$.

Here we allow different degrees of smoothness for different components of the time-varying coefficient. The smoothness condition on $\lambda_\tau(t)$ is suited for the use of smoothing splines. The number of measurements n_i made on the ith subject is considered random, reflecting sparse and irregular designs, and is required to satisfy the following assumption

Assumption 2. $\{n_i, i = 1, \ldots, m\}$ are i.i.d. rv's with finite expected value, and independent of all other random variables.

We use the following notation. Let $\beta(\lambda, t) = \text{argmin}_b E\{\rho_\tau(\Lambda(Y(t), t; \tau) - X(t)^T b)\}$, and $\mathcal{B}$ be a subset of R^p containing $\beta(\lambda, t)$ for all $\lambda \in \Omega_\tau$ and $t \in [0, 1]$. Let $\dot{\beta}_t(\lambda, t)$ and $\dot{\beta}_\lambda(\lambda, t)$ denote the first derivatives of $\beta(\lambda, t)$ with respect to t and λ, respectively. Let $F(\cdot|t, x, \lambda)$ denote the conditional distribution function of $\Lambda(Y(t), t; \tau) - X(t)^T\beta(\lambda, t)$ given $X(t) = x$, and $f(\cdot|t, x, \lambda)$ the corresponding conditional density.

Assumption 3. $\Omega_\tau \bigotimes \mathcal{B}$ is a compact set of $\mathcal{R} \bigotimes \mathcal{R}^p$. $\lambda_\tau(t)$ is an interior point of Ω_τ for each t.

Assumption 4.

(i) There is at least one component of $X(T)$ whose conditional distribution given $T = t$ is absolutely continuous with respect to Lebesgue measure for all $t \in [0, 1]$.

(ii) $E(X(T)X(T)^T|T = t)$ is positive definite for every $t \in [0, 1]$.

Assumption 5. The distribution of T is absolutely continuous with a density function bounded away from zero on $[0, 1]$.

Assumption 6.

(i) The support of $X(t)$, $\mathcal{X}$, is bounded uniformly in t.

(ii) There exists a constant B such that, for all x_1 and x_2 in $\mathcal{X}$, $|G(x_1|t) - G(x_2|t)| \leq B\|x_1 - x_2\|$, where $G(\cdot|t)$ denotes the distribution function of $X(T)$ at a fixed t.

Assumption 7.

(i) There exists an integrable function $M(x, t)$ such that $f(u; t, x, \lambda) \leq M(x, t)$, uniformly in $\lambda \in \Omega_\tau$ for all $x \in \mathcal{X}$ and all $t \in [0, 1]$.

(ii) $f(0; t, x, \lambda) > 0$ for all x, all $x \in \mathcal{X}$, $t \in [0, 1]$, and $\lambda \in \Omega_\tau$.

(iii) There exists a positive definite matrix D_λ such that $n^{-1}\sum_{ij} E[X_{ij}X_{ij}^T f(0; T_{ij}, X_{ij}, \lambda)]$ converges to D_λ in probability; the smallest eigenvalue of D_λ is bounded above from zero uniformly in $\lambda \in \Omega_\tau$.

Assumption 8. The first derivatives of $\beta(\lambda, t)$ are bounded for all $t \in [0, 1]$ and all $\lambda \in \Omega_\tau$.

Assumption 9. There exists an integrable function $L(x,t)$ such that, for any $\lambda_1, \lambda_2 \in \Omega_{\lambda_\tau}$

$$|F(0;t_1,x,\lambda_1) - F(0;t_2,x,\lambda_2)| \le L(x,t)|\lambda_1 - \lambda_2| \quad \text{for all} \quad x \in \mathcal{X} \quad \text{and} \quad t \in [0,1].$$

Assumption 10. Let $\zeta = (\lambda, \beta, t)$ and $\zeta_1 = (\lambda_1, \beta_1, t_1)$, and take

$$\phi(x,\zeta) = E\Big\{\mathrm{I}\{X_i(T) \le x\}[\tau - F(X_i(T)^T(\beta - \beta(\lambda,t)); t, X_i(T), \lambda)]|T = t\Big\}.$$

There exists a constant $M(x)$ such that, for any two points ζ and ζ_1 in $\Omega_{\lambda_\tau} \bigotimes \mathcal{B} \bigotimes [0,1]$, $|\phi(x,\zeta) - \phi(x,\zeta_1)| \le M(x)|\zeta - \zeta_1|$ for all $x \in \mathcal{X}$, where $|\cdot|$ denotes the sup norm.

Assumptions 3 and 4 are sufficient conditions for identifiability of $\lambda_\tau(t)$ and $\beta_\tau(t)$ at all $t \in [0,1]$. Assumptions 5-10 are sufficient to derive the uniform consistency of $\hat{\lambda}_n(t)$ and the consistency of $\hat{\beta}_{n,k}(t)$ in the sense of (2.10), and therefore the consistency of the estimated conditional quantiles of $Y(t)$ given $X(t)$ at a fixed t. This set of assumptions is stronger than needed, but it simplifies technical details without losing much generality. We also point out that our arguments are made without reference to homoscedasticity, thus the covariates X_{ijk}, $k = 1, \ldots, p$, may correlate with the error term e_{ij}.

3. Application

We illustrate the proposed method using a HIV dataset collected by the AIDS Clinical Trials Group. In this study, 517 HIV-1 infected patients were randomly assigned to three treatments for 120 weeks, and their CD4 cell counts were monitored at weeks 4, 8, and every 8 weeks thereafter. More details about the dataset can been found in Park and Wu (2005). Here we explore the clinical applications of model (1.1) in two aspects: construction of a prediction band for the longitudinal response variable, and interpretation of covariate effects. Due to big drop-out rates after 100 weeks, only the data collected during the first 100 weeks are included in the analysis.

Model estimation. A set of conditional quantile functions of a longitudinal response variable provides a summary of the conditional distribution of the response variable conditional on the covariates. By studying a set of quantile functions, we can learn about the location, skewness and other aspects of the conditional distribution. Here we study what effects treatments have on the shape of the response distribution, and how they differ across different locations of the response distribution from lower percentiles to upper percentiles. For this purpose, we focused on the median, 10th and 90th quantiles of the response variable, conditional on treatment and initial disease severity. Let $Y_i(t)$ be the CD4

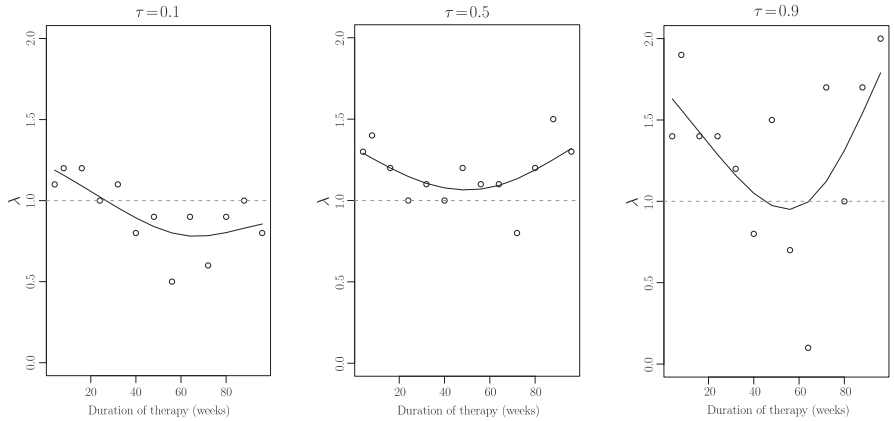

Figure 1. The estimated power function $\lambda_\tau(t)$. From left to right the quantile levels are 0.1, 0.5 and 0.9, respectively.

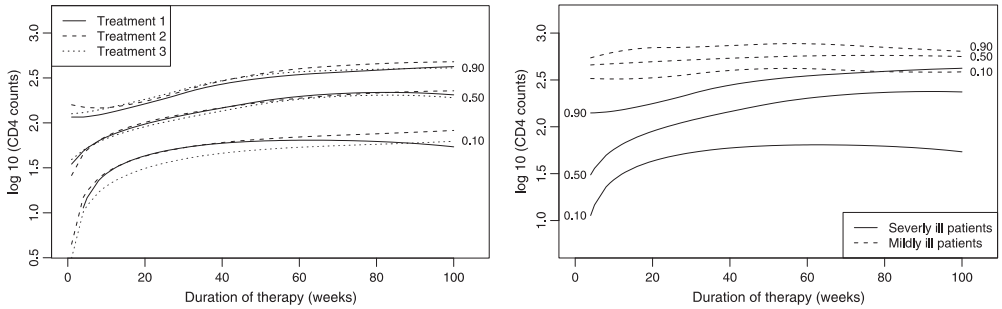

Figure 2. Estimated conditional 10th, 50th and 90th quantile functions. Left panel: conditional quantiles of CD4 cell counts of severely-ill patients under three different treatment; Right panel: conditional quantiles of CD4 cell counts of severely-ill and mildly-ill patients under Treatment 1.

cell counts of the i-th patient at time t, and $Y_i(0)$ denote the baseline CD4 cell counts at time $t = 0$. We modeled $Y_i(t)$ by

$$\Lambda(Y_i(t), t; \tau) = \beta_{\tau,0}(t) + \beta_{\tau,1}(t)Y_i(0) + \beta_{\tau,2}(t)Z_{1,i} + \beta_{\tau,3}(t)Z_{2,i} + \epsilon_i(t; \tau), \quad (3.1)$$

where t is the therapy duration in weeks, $Z_{1,i}$ and $Z_{2,i}$ are binary indicators for the first and the second treatment groups, respectively. The error term $\epsilon_i(t; \tau)$ has zero conditional τ-th quantile given the covariates.

We partitioned the time interval by the mid-points between the scheduled follow-up times. The raw estimates of the transformation functions at those mid-points are displayed in Figure 1 with the solid lines representing the smoothed estimates. At all the quantile levels, the smoothed transformation curves present

a quadratic pattern: they decrease after the therapy onset, and then rise after 40−60 weeks. To gain some confidence in the need for power transformations, we applied the linearity test proposed by Mu and He (2007) which tests the hypothesis $H_0 : \lambda_\tau(t) = 1$ at a fixed time point. Even though the method does not strictly apply to correlated data, we applied it to each subinterval partitioned by the scheduled follow-up times. The p-values during the early stages receiving treatment were significant.

To demonstrate treatment effects, we display in the left panel of Figure 2 the estimated median, 10th, and 90th quantile curves of CD4 Cell counts for severely-ill patients whose baseline CD4 cell counts are as low as 10, but under the three treatments. The coefficient functions $\beta_{\tau,k}(t)$'s were estimated by normalized B-splines with the number of knots chosen by the model selection criterion defined in (2.8). For illustration purposes only, the estimated quantile curves are plotted on $\log 10$ scale. As suggested by the left panel of Figure 2, all three treatments elevate severely-ill patients' CD4 cell counts, especially during the first 40 weeks but with non-differential their effects. The right panel displays how the effects of Treatment 1 change from the 10th percentile to the 90th percentile for patients on two severity levels: the solid ones represent the conditional quantiles based on the severely-ill patients; the dotted curves are based on the mildly-ill patients whose initial CD4 cell counts are 398. The three quantiles of the conditional distribution of the CD4 cell counts of severely-ill patients increase rapidly at the beginning, and continue to improve at a gradually slower rate. Moreover, the conditional distribution of CD4 cell counts of this group is skewed to the lower end by comparing the spacings between the 10th and 90th quantile from the median. Had we used the Gaussian approach to construct the prediction band, the quantiles may have been estimated with bias. As compared to the severely-ill patients, the mildly-ill patients respond to the treatment less actively. Their CD4 cell counts are more stable over the treatment duration, and the distribution is more symmetric around the median.

Comparison to alternative methods. We consider two alternative methods as follows.

1. *The simple varying-coefficient quantile regression model* (Kim (2007)), which assumes

$$Y_i(t) = \beta_{\tau,0}(t) + \beta_{\tau,1}(t)Y_i(0) + \beta_{\tau,2}(t)Z_{1,i} + \beta_{\tau,3}(t)Z_{2,i} + \xi_i(t;\tau), \qquad (3.2)$$

   where the error term $\xi_i(t;\tau)$ has zero conditional τ-th quantile given the covariates. One may approximate the coefficient functions using normalized quadratic B-splines with knots selected by (2.8). We refer to Kim (2007) for technical details.

2. *The extended LMS model*, which assumes the response $Y(t)$ can be transformed to a Gamma distribution by

$$W(t) = \Big[\frac{Y(t)}{\mu}\Big]^{\lambda} \sim Gamma(\text{mean} = 1, \text{variance} = \lambda^2\sigma^2), \tag{3.3}$$

where λ, μ, and σ are power, location, and scale parameters all belonging to a family of semi-parametric functions of the form

$$g = s(t) + \beta_1 Y_i(0) + \beta_2 Z_{1,i} + \beta_3 Z_{2,i}, \quad \beta_i \in \mathbb{R},$$

where $s(t)$ is a smooth function of time t.

The parameter functions λ, μ, and σ can be estimated via maximizing a penalized likelihood. The τth quantile of $Y(t)$ at time t can then be constructed from estimated parameter functions and the τth quantile of the *Gamma* distribution. Model (3.3) was proposed by Yee (2004) as an extension of the LMS method proposed by Cole and Green (1992).

The simple varying-coefficient quantile model becomes a special case of Model (1.1) by taking $\lambda(t) = 1$. A comparison between the two models demonstrates what a time-dependent transformation adds to the understanding of the data. The extended LMS model estimates the conditional quantile functions based on time-dependent power transformation, but relies on a parametric distribution assumption, and uses the same transformation at all the quantiles. The left panel of Figure 3 displays the three sets of quantile curves for severely-ill patients based on Models (1.1) and (3.2), respectively. The solid lines denote the quantile estimates from Model (1.1), the dashed lines from the untransformed counterpart, and the dotted line from the extended LMS model. The three models agree with each other during late stages of the treatment, however they differ in the early stages of therapy (0−30 weeks). The quantile estimates from the simple, varying-coefficient quantile model are higher than those from its transformed counterpart at all the three quantile levels.

To evaluate model fits and investigate the discrepancies among the models that surface in Figure 3, we performed a five-fold cross-validation based on a score function defined as the standardized difference between the proportion of negative residuals and the quantile level τ. That is,

$$\hat{d} = \frac{\sum_i \sum_j \mathrm{I}\{y_{ij} - \hat{Q}_{y_{ij} \le 0}(\tau | t_{i,j}, x_{i,j})\}}{\sqrt{n\tau(1-\tau)}} - \tau.$$

We expect the difference, $\hat{d}$, to be close to 0 for a good model fit for any subgroup. To check the model fit during the early stage of treatment for severely-ill

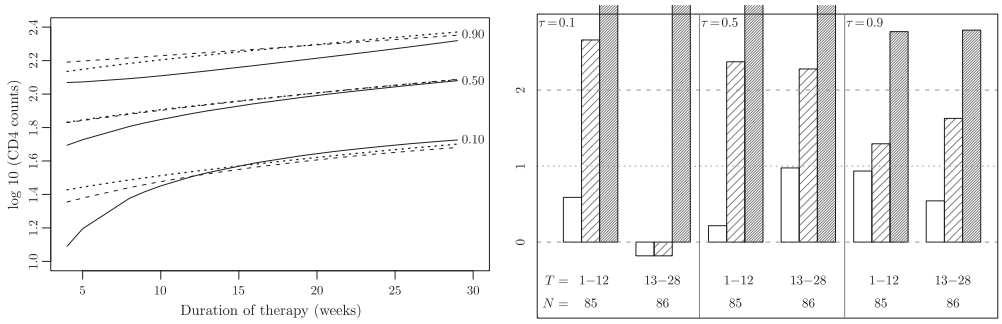

Figure 3. Comparison of the simple varying-coefficient quantile regression model and the extended LMS model. Left panel: the solid lines are estimated quantiles of the CD4 cell counts of severely ill patients at $\tau = 0.1$, 0.5 and 0.9, respectively; the dashed lines are estimated quantiles from the simple varying-coefficient quantile regression model, and the dotted lines are those from the extended LMS method. Right panel: the light bars are 5-fold cross-validation scores $\hat{d}_{cv}$ from Model (1.1), the gray ones are those from the simple varying-coefficient quantile regression model, and the dark gray ones are those from the extended LMS model. T is the treatment duration in weeks, and N is the number of measurements in each time intervals.

patients, we selected subjects whose baseline CD4 cell counts were 10 ± 5. Let Γ_1 denote the set of measurements taken in weeks 1−12 and Γ_2 denote the set of measurements taken in weeks 13−28 on those severely-ill patients. We calculated cross-validation scores based on Γ_1 and Γ_2, respectively:

$$\hat{d}_{cv}(\Gamma_m) = \frac{\sum_{k=1}^{5} \sum_{i \in S_k} \sum_j \mathrm{I}\{(i,j) \in \Gamma_m\} \mathrm{I}\{Y_{ij} \le \hat{Q}_{y_{ij}}^{-S_k}(\tau | t_{ij}, x_{ij})\}}{\sqrt{\#(\Gamma_m)\tau(1-\tau)}} - \tau, \quad m = 1, 2,$$

where $\#(\Gamma_m)$ is the number of measurements contained in Γ_m, S_k, $k = 1, \ldots, 5$, are sets of indices for the randomly partitioned cross-validation sets, and $\hat{Q}^{-S_k}$ denote the quantile estimates using the full data excluding measurements in S_k. The cross-validation scores based on Γ_1 and Γ_2 for the three competing models are displayed in Figure 3. The light bars represent the cross-validation scores based on the fit of Model (1.1), the gray ones are scores based on the fit of the simple varying-coefficient quantile regression model, and the dark gray ones are those based on the extended LMS model. As seen in Figure 3, the cross-validation scores from Model (1.1) are uniformly better than those of the other two models. Both the untransformed quantile regression model and the extended LMS model apparently over-estimated the CD4 cell counts of the severely ill patients during the early stage of treatment; still, the untransformed model provides a reasonable

fit. We also evaluated the cross-validation scores over other time intervals, and the latter two models were comparable with each other.

Interpretation of covariate effects. Unlike the simple varying-coefficient quantile regression model, the linear coefficients $\beta_{\tau,k}$ in Model (1.1) do not provide direct interpretation on covariate effects. However one can examine marginal covariate effects defined as the first derivative of the quantile function with respect to the covariate. Marginal covariate effect measures the change of the τth conditional quantile of the response variable due to a unit increase in the covariate. Because of the use of transformations, the marginal covariate effect is not only a function of time t, but also a function of the covariate itself. This feature yields interesting findings of the association between the covariate and the response variable. Here the quantile function is

$$\hat{Q}_{Y_{ij}}(\tau|t_{ij}, x_{ij}) = S(\hat{\lambda}_n(t_{ij}), \hat{q}(t_{ij}, x_{ij})), \tag{3.4}$$

where $S(\lambda, u) = (\lambda \cdot u + 1)^{1/\lambda}$ ($= \exp(u)$ if $\lambda = 0$) denotes the inverse power transformation function, and $\hat{q}(t_{ij}, x_{ij}) = \hat{\beta}_{n,1}(t_{ij})x_{ij1} + \cdots + \hat{\beta}_{n,p}(t_{ij})x_{ijp}$. Based on (3.4), the τth conditional quantile of CD4 cell counts at time t under Treatment 1 can be estimated by $S(\hat{\lambda}_n(t), q(t, x(1))$, where $\hat{q}(t, x(1)) = \hat{\beta}_{n,0}(t) + \hat{\beta}_{n,1}(t)Y(0) + \hat{\beta}_{n,2}(t)$. As a result, the estimated marginal effect of the baseline CD4 cell counts under Model (3.1) at time t is

$$\hat{\beta}_{n,1}(t) \cdot \left\{ \hat{\lambda}_n(t) \cdot \left(\hat{\beta}_{n,0}(t) + \hat{\beta}_{n,1}(t)Y(0) + \hat{\beta}_{n,2}(t) \right) + 1 \right\}^{[1-\hat{\lambda}_n(t)]/[\hat{\lambda}_n(t)]}.$$

Note that the marginal effect of a covariate changes with the covariate. Figure 4 displays the estimated marginal effects of baseline CD4 cell counts (black lines) at weeks 4 and 12 at the three quantile levels. We also superimpose the corresponding covariate effects (gray horizontal lines) from the simple varying-coefficient quantile regression model. As suggested by Figure 4, at week 4, the marginal baseline CD4 effect on all the three quantiles starts high at small baseline CD4 cell counts (corresponding to those severely-ill patients) and decreases quickly as the baseline CD4 cell counts increase. The marginal baseline CD4 effect at week 12 follows a similar pattern but decreases at a much slower rate. The simple varying-coefficient quantile regression model however does not capture this nonlinear pattern, and underestimates the baseline CD4 effects at low baseline CD4 cell counts. This partly explains the lack-of-fit for the severely-ill patients when a simple varying-coefficient model is used.

In this example, Model (1.1) demonstrates its flexibility to capture the nonlinearity and reduce the bias from that of more parametric models.

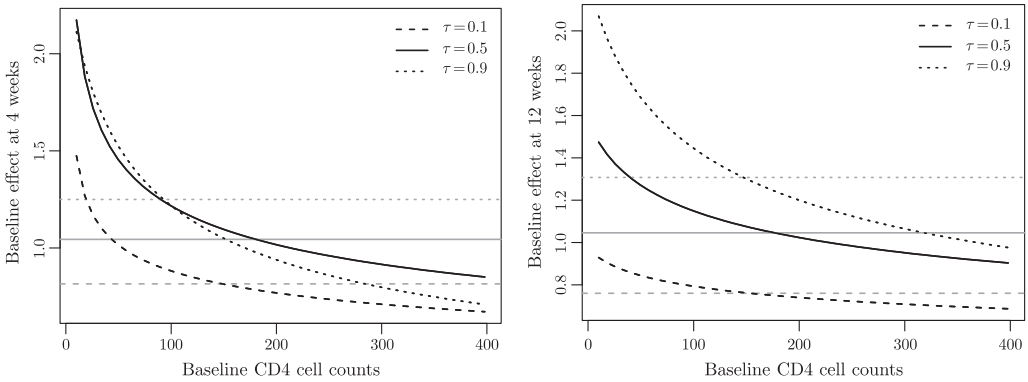

Figure 4. Comparison of baseline CD4 effects between Model (1.1) and the simple varying-coefficient quantile regression model at week 4 (left panel) and week 12 (right panel), respectively. The black curves denote the baseline CD4 effects from Model (1.1), and the grey horizontal lines from the simple varying-coefficient quantile regression model. Solid, dotted, and dashed lines represent the median, the 10th, and the 90th quantiles, respectively.

4. Monte Carlo Simulation

The aim of this section is to investigate the finite sample performance of the two-step procedure, and study the sensitivity of the performance to the choice of k_n. In the simulation, the R function *smooth.spline* was called repeatedly where the GCV option is applied in selecting the smoothing parameter for the transformation function. The BIC type criterion introduced in (2.8) was used to select knots for a quadratic spline estimator of the coefficient functions.

Throughout the simulation, we generated data from the model designed as follows.

$$\Lambda(Y_{ij}, t_{ij}; \lambda_\tau) = \beta_0(t_i j) + \beta_1(t_{ij}) X_{ij1} + \beta_2(t_{ij}) X_{ij2} + \varepsilon_{ij},\ i = 1, \ldots, m;\ j = 1, \ldots, n_i. \tag{4.1}$$

We let the time interval be [0, 1] and chose 40 time points equidistant over [0, 1]. We considered $m = 200, 300$ and 400, respectively. We drew a random sample of size m from the distribution Unif(0, 30), and took the ceilings of this random sample to be $\{n_i\}_{i=1}^m$, so that each n_i was at least 1. The transformation function was taken to be $\lambda_\tau(t) = (t-1)^2 + 1/2$. Three coefficient functions were chosen as $\beta_0(t) = 4 + (t - 1/2)^2$, $\beta_1(t) = 3 + t$, and $\beta_2(t) = \exp(t)$. Three covariates were chosen: $X_0(t) = 1$; $X_1(t) = t^2 + |Z|$, where Z is a standard normal random variable; $X_2(t)$ with a time-invariant uniform distribution on the interval [0,4]. The errors were sampled from a stationary Gaussian process with a decaying exponential covariance function

$$\mathrm{Cov}(\varepsilon_{i_1}(t_1), \varepsilon_{i_2}(t_2)) = \exp(-2 \cdot |t_1 - t_2|)) \quad \text{if } i_1 = i_2;\ 0, \text{ otherwise.}$$

For each choice of m, we sampled 100 data sets from model (4.1) and fit them by the two-step procedure. We fixed $\tau = 0.5$, and considered $k_n = 10, 15,$ and 20. We evaluated the relative efficiency of a fit using the two-step procedure compared to using a simple varying-coefficient quantile regression model. For that purpose, we computed a MSE ratio at each replication. Let $\hat{Q}_{y_{ij}}(\tau|t_{ij}, x_{ij})$ denote the fitted quantiles based on estimation of model (1.1) and $\tilde{Q}_{y_{ij}}(\tau|t_{ij}, x_{ij})$ denote the fitted quantiles based on estimation of model (1.1) setting $\lambda_\tau(t) = 1$. At the rth replication, we computed the mean squared error of $\hat{Q}_{y_{ij}}(\tau|t_{ij}, x_{ij})$ from the true quantile function as

$$\mathrm{MSE}_r^1 = \frac{1}{n}\sum_{i,j}\left(Q_{y_{ij}}(\tau|t_{ij}, x_{ij}) - \hat{Q}_{y_{ij}}(\tau|t_{ij}, x_{ij})\right)^2. \tag{4.2}$$

Similarly we denote the mean squared error of $\tilde{Q}_{y_{ij}}(\tau|t_{ij}, x_{ij})$ by MSE_r^0. The mean of the ratios $\mathrm{MSE}_r^1/\mathrm{MSE}_r^0$, $r = 1, \ldots, 100$, measures the relative efficiency of model (1.1) versus a simple, varying-coefficient quantile regression model. Some summary statistics of the simulation results are presented in Table 1.

First we examine the simulation results when $m = 200$. Notice that the mean ratios are larger than one if $k_n = 10$ or 15, which suggests the sensitivity of the two-step procedure to the choice of k_n. We investigated this issue and it turned out that the mean ratios are inflated by less than 10% of the runs where there is undersmoothing of the raw estimates of the transformation function. When we increase m to 300 or 400, the mean ratios fall below 1 regardless of the choice of k_n. Looking at other summary statistics of the ratios confirms that the two-step procedure can effectively uncover the true underlying feature of a data set, and thus improve efficiency compared to using a simple, varying-coefficient quantile regression model.

Our simulation studies suggest that GCV is a reasonable criterion under our setting though it is not perfect (it undersmoothes 10% of the time in our simulation). The small sample size of raw estimates, the heteroscedasticity in the raw estimates, as well as the correlation among the raw estimates, all could lead to a failure of GCV.

5. Concluding Remarks

We propose a flexible transformed varying-coefficient model for longitudinal data based on modelling conditional quantiles of the longitudinal response variable. Time-dependent power transformations are used to achieve linearity, while the use of quantile regression relaxes the parametric distributional assumptions. We have shown through a data example and simulation studies that the proposed method can help estimate the quantiles when the sample size is sufficiently large. In this case, the transformation offers an opportunity to reduce

Table 1. Summary statistics of ratios of mean squared errors of the fitted quantiles from the proposed model to those from a simple, varying-coefficient quantile regression model. TM 1 is calculated as the mean of the remainder after eliminating 10% of the values at the right end of the ordered sample. TM 2 is calculated as the mean of the remainder after eliminating 5% of the values at the right end of the ordered sample.

	k_n	Mean	TM 1	TM 2	Median	75% Percentile	90% Percentile
	10	5.106	0.252	0.307	0.217	0.413	0.954
m=200	15	0.719	0.194	0.237	0.187	0.260	0.638
	20	1.623	0.246	0.524	0.195	0.316	1.766
	10	0.524	0.199	0.261	0.141	0.262	1.09
m=300	15	0.241	0.119	0.140	0.117	0.171	0.354
	20	0.969	0.277	0.480	0.122	0.213	2.85
	10	0.974	0.190	0.354	0.125	0.276	1.445
m=400	15	0.568	0.119	0.154	0.096	0.183	0.410
	20	0.446	0.108	0.133	0.097	0.167	0.301

bias from more parametric models. Model checking and diagnostic tools help one decide whether it is worth the efforts in specific applications. We also notice that GCV may sometimes undersmooth. We therefore recommend, in practice, doing model diagnostics to validate the use of a statistical model, such as the use of cross-validation in Section 3. A more stable method for choosing the smoothing parameter in the current setting is desired, but is delayed to future work.

Acknowledgements

The authors are grateful to two referees, associate editor, and consulting editor, whose comments led to a much improved manuscript. This work was partly supported by the National Science Foundation Awards DMS-0504972.

References

Cai, Z. and Xu, X. (2008). Nonparametric quantile estimations for dynamic smooth coefficient models. *J. Amer. Statist. Assoc.* **103**, 1595-1608.

Chiang, C.-T., Rice, J. and Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variable. *J. Amer. Statist. Assoc.* **96**, 605-619.

Cole, T. J. and Green, P. J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. *Statist. Medicine* **11**, 1305-1319.

Fan, J. and Zhang, J. T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *J. Roy. Statist. Soc. Ser. B* **62**, 303-322.

Friedman and Silverman. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31**, 3-39.

Hastie, T. J. and Tibshirani, R. J. (1993). Varying-coefficient models (with discussion). *J. Roy. Statist. Soc. Ser. B* **55**, 757-796.

He, X. and Shi, P. (1996). Bivariate Tensor-Product B-Splines in a Partly Linear Model. *J. Multivariate Anal.* **58**, 162-181.

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

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.

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

Lipsitz, S., Fitzmaurice, G., Molenberghs, G. and Zhao, L. P. (1997). Quantile regression methods for longitudinal data with drop-outs: application to CD4 cell counts of patients infected with the human immunodeficiency virus. *Appl. Statist.* **46**, 463-476.

Mu, Y. (2005). Power transformation towards linear or partially linear quantile regression models. Ph.D. thesis, The University of Illinois at Urbana-Champaign.

Mu, Y. and He, X. (2007). Power transformation towards a linear regression quantile. *J. Amer. Statist. Assoc.* **102**, 269-279.

Nychika, D. (1995). Splines as local smoothers. *Ann. Statist.* **23**, 1175-1197.

Park, J. G. and Wu, H. (2005). Backfitting and local likelihood methods for nonparametric mixed-effects models with longitudinal data. *J. Statist. Plann. Inference* **136**, 3760-3782.

Portnoy, S. and Mizera, I. (1998). Comment on "instability of least squares, least absolute deviation, and least median of squares linear regression" by S. P. Ellis. *Statist. Sinica* **13**, 344-347.

Wang, Y. (1998). Smoothing spline models with correlated random errors. *J. Amer. Statist. Assoc.* **93**, 341-348.

Wei, Y. and He, X. (2006). Conditional growth charts (with discussion). *Ann. Statist.* **34**, 2069-2097.

Wei, Y., Pere, A., Koenker, R. and He, X. (2005). Quantile regression methods for reference growth charts. *Statist. Medicine* **25**, 1369-1382.

Wu, C. O., Yu, K. F. and Chiang, C.-T. (2000). A two-step smoothing method for varying coefficient models with repeated measurements. *Ann. Inst. Statist. Math.* **52**, 519-543.

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

Yee, T. W. and Wild, C. J. (1996). Vector generalized additive models. *J. Roy. Statist. Soc. Ser. B* **58**, 481-493.

Yee, T. W. (2004). Quantile regression via vector generalized additive models. *Statist. Medicine* **23**, 2295-2315.

Department of Mathematics and Statistics, Portland State University, PO Box 751, Portland, Oregon 97207-0751, U.S.A.

E-mail: yunmingm@pdx.edu

Department of Biostatistics, Columbia University, New York, NY 10032. U.S.A.

E-mail: yw2148@columbia.edu

(Received March 2007; accepted February 2008)