The Annals of Applied Statistics
2015, Vol. 9, No. 2, 597–620
DOI: 10.1214/15-AOAS811

REGRESSION BASED PRINCIPAL COMPONENT ANALYSIS FOR SPARSE FUNCTIONAL DATA WITH APPLICATIONS TO SCREENING GROWTH PATHS

BY WENFEI ZHANG AND YING WEI

Columbia University

Growth charts are widely used in pediatric care for assessing childhood body size measurements (e.g., height or weight). The existing growth charts screen one body size at a single given age. However, when a child has multiple measures over time and exhibits a growth path, how to assess those measures jointly in a rigorous and quantitative way remains largely undeveloped in the literature. In this paper, we develop a new method to construct growth charts for growth paths. A new estimation algorithm using alternating regressions is developed to obtain principal component representations of growth paths (sparse functional data). The new algorithm does not rely on strong distribution assumptions and is computationally robust and easily incorporates subject level covariates, such as parental information. Simulation studies are conducted to investigate the performance of our proposed method, including comparisons to existing methods. When the proposed method is applied to monitor the puberty growth among a group of Finnish teens, it yields interesting insights.

1. Introduction. In pediatric practice, height, weight and other body size measurements are frequently examined for infants, children and adolescents in order to ensure their healthy growth. The most commonly used tools are growth charts, also known as reference centile charts. The fundamental purpose of growth charts is to identify percentile ranks of individuals with respect to their corresponding reference populations, and to screen out subjects with extreme ranks, either too high or too low, for further medical investigations. The conventional growth charts consist of a series of percentile curves for a certain measurement over ages. Those percentile curves are estimated from a reference population using penalized likelihood methods introduced in Cole (1988) and Cole and Green (1992). They are used to identify individual percentile ranks at specific ages. Lately, several methods, including Thompson and Fatti (1997), Scheike, Zhang and Juul (1999), Wei et al. (2006) and Chen and Müller (2012), were proposed to further incorporate prior information and subject level covariates into growth charts. In these methods, the reference percentiles are estimated by conditioning on not only target

Received May 2014; revised January 2015.
Key words and phrases. Growth charts, sparse functional data, longitudinal data, principal component analysis.

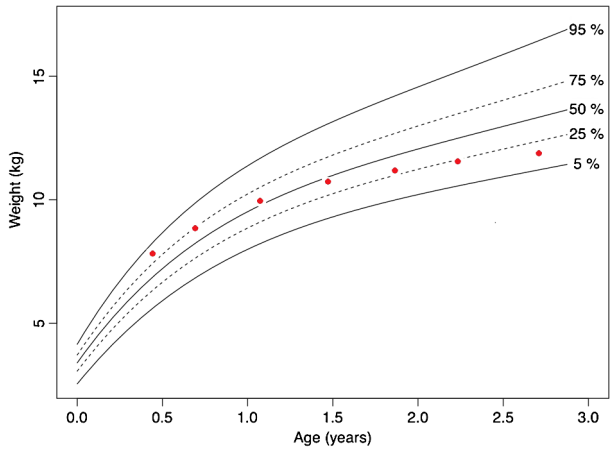


FIG. 1. *An example of an abnormal growth pattern. The dots represent the growth path for a subject. The curves are the percentile curves at quantile levels* 0.05, 0.25, 0.50, 0.75 *and* 0.95. *The x axis represents ages. The y axis represents weights.*

ages but also prior measurements and other important variables, such as prognostic and parental information. The resulting growth charts are hence called conditional growth charts. Thompson and Fatti (1997) assumed a multivariate normal distribution for the measurements and the covariates at all time points and used the maximum likelihood estimator for the mean and variance functions. Scheike, Zhang and Juul (1999) considered a longitudinal regression model accounting for the previous measurement adjacent to the current measurement and the duration in between. To avoid a particular distributional assumption, Wei et al. (2006) proposed a semi-parametric quantile regression model to construct conditional growth charts.

Both conventional and conditional growth charts screen only one single measurement at a time. However, due to common clinical practice, each individual has its measurements collected longitudinally and exhibits a growth path over time. A growth path may not be normal even if each of its measurements is within the normal ranges of both conventional and conditional growth charts. For example, as shown in Figure 1, this subject starts at the 90th percentile in weight at the age of 0.5, and gradually declines to the 15th percentile around the age of 2.5. Although such a slow decline in the growth path should be alerting, it cannot be recognized by conventional growth charts, because all its measurements are within the normal ranges. It cannot be detected by conditional growth charts either, since the changes from the preceding measurements are not large enough.

Therefore, screening entire growth paths may bring new insights into growth screening. However, existing screening methods for growth paths are mostly empirical, relying heavily on personal experiences of medical providers [Legler and Rose (1998)]. Rigorous quantitative screening methods for entire growth paths

remain largely undeveloped. Hence, in this paper, we propose a new statistical method to construct growth charts that enable the screening of entire growth paths.

Growth charts are estimated from a reference growth data set, which is collected from a representative sample in a target population, and consists of longitudinal body size measurements. Most reference growth data share the following characteristics. First, each growth path is only observed at sparse and irregularly spaced time points with possible measurement errors. Therefore, statistical tools developed for multivariate and functional data are directly applicable, as the former requires a fixed measurement schedule and the latter requires densely observed data on each growth path. Second, the distributions of body size measurements are unlikely to follow certain parametric distributions. Therefore, likelihood based parametric approaches are often undesirable in such applications.

Considering the characteristics of reference growth data, we develop a two-step procedure for identifying percentile ranks of growth paths. In the first step, we propose a novel regression based principal component analysis (PCA) algorithm that is tailored specifically for reference growth data. In the second step, we construct the multivariate quantile contours of the resulting component scores, which can be used to identify percentile ranks of growth paths. The proposed PCA algorithm can also incorporate covariates, which in turn enables the screening of growth paths conditioned on individual characteristics.

The rest of this paper is organized into the following structure: In Section 2 we elaborate on the proposed screening method, including the general model settings and notation in Section 2.1, the introduction of the proposed regression based principal component analysis in Section 2.2, the construction of growth charts for screening growth paths in Section 2.3, and the extension of incorporating covariates in Section 2.4. In Section 3 we provide examples of applying the proposed method in the field of pediatrics. In Section 4 we present the numerical investigation of our method. In Section 5 we include discussions and conclusions on the important findings.

2. Methods.

2.1. *Settings and notation.* A reference growth data set consists of N subjects and their longitudinal measurements $\{Y_{ij}, T_{ij}\}_{i=1,\ldots,N,\, j=1,\ldots,m_i}$. Here m_i is the number of measurements for the ith subject, and Y_{ij} is the jth observation for the ith subject measured at the time of T_{ij}, $T_{ij} \in \mathcal{T}$. We assume that each longitudinal growth path is observed from the following model:

$$Y_{ij} = Y_i(T_{ij}) + \varepsilon_{ij}, \qquad T_{ij} \in \mathcal{T}, \tag{1}$$

where $Y_i(t)$'s are the underlying growth paths, ε_{ij}'s, and independent of $Y_i(t)$'s, are i.i.d. random errors with mean zero and constant variance σ^2. ε_{ij} can be viewed as the measurement error associated with Y_{ij}, and we implicitly assume that measurement errors do not depend on magnitudes of measurements and measurement

times. Such assumptions are reasonable for reference growth data. For example, the weight measurement error due to a weight scale is usually related neither to the weight itself nor to the time when the weight is taken.

By the Karhunen–Loève theorem in Loeve (1978), the true growth paths $Y_i(t)$, if smooth and continuous, can be written as

$$Y_i(t) = U(t) + \sum_{k=1}^{\infty} r_{ik}\phi_k(t), \tag{2}$$

where $U(t) = E\{Y_i(t)\}$ is the population mean function, $\phi_k(t)$'s are principal component functions, which are continuous pair-wise orthogonal functions on $\mathcal{T}$ with $\int_{\mathcal{T}} \phi_k(t)^2\,dt = 1$, and r_{ik}'s are principal component scores, which are uncorrelated random variables with mean 0 and variance λ_k, where $\lambda_1 \geq \lambda_2, \ldots$. This decomposition provides the basis of PCA for functional data.

We further assume $Y_i(t)$ can be well approximated by the first K principal component functions, that is, $Y_i(t) \approx U(t) + \sum_{k=1}^{K} r_{ik}\phi_k(t)$. This approximation is biologically plausible, since the biological growth process is mainly driven by several growth hormones, as mentioned in Zhang (2012). As each growth hormone determines a particular growth pattern, the observed growth path is the result of their joint actions. Therefore, with the kth component function ϕ_k representing a certain growth pattern, the component score r_{ik} measures the extent to which $\phi_k(t)$ contributes to the individual growth path $Y_i(t)$. The biological meaning of component functions $\phi_k(t)$ and scores r_{ik} is also exemplified in Section 3.1. This way, the distribution of the growth paths, $Y_i(t)$'s, are fully determined by their component scores. Consequently, the growth charts for $Y_i(t)$ can be constructed based on the joint distribution of the first K component scores. To estimate the component functions of $Y_i(t)$ from the reference growth data, we proposed a regression based PCA algorithm in Section 2.2.

The following notation will be used to illustrate our proposed method: $L^2(\mathcal{T})$ is the set of square integrable functions defined on the time interval $\mathcal{T}$. Denote $\|\cdot\|^2$ as the L^2 norm for functions in $L^2(\mathcal{T})$, that is, $\|f\|^2 \hat{=} \int_{\mathcal{T}}\{f(t)\}^2\,dt$, $\forall f(t) \in L^2(\mathcal{T})$. The inner product of two functions $f_1(t)$ and $f_2(t)$ in $L^2(\mathcal{T})$ is defined as $\langle f_1, f_2\rangle \hat{=} \int_{\mathcal{T}} f_1(t)f_2(t)\,dt$. When $\langle f_1, f_2\rangle = 0$, we say that $f_1(t)$ and $f_2(t)$ are orthogonal to each other, denoted as $f_1 \perp f_2$.

2.2. *Regression based principal component analysis for growth data.* Reference growth data can be considered as sparse functional data due to the sparse and irregular data structure. There exists a few PCA methods for sparse functional data, including Yao, Müller and Wang (2005), James, Hastie and Sugar (2000) and Peng and Paul (2009). Yao, Müller and Wang (2005) involved the estimation of high-dimensional covariance matrices, as well as their inverses, which may not be computationally stable. James, Hastie and Sugar (2000) provided a stable maximum likelihood estimation (MLE) algorithm under the assumption of Gaussian

process. Peng and Paul (2009) implemented the same model from James, Hastie and Sugar (2000) using an improved fitting procedure. However, the distribution assumption of the MLE methods may not be satisfied by reference growth data. In this section, we propose a regression based PCA algorithm which is computationally stable, not relying on strong distribution assumptions, and easily incorporates covariates. Without loss of generality, and to simplify the notation, we assume in this section that the population mean $U(t)$ in (2) is 0. For nonzero $U(t)$, we can get its nonparametric estimation and subtract it from $Y_i(t)$. The algorithm can be applied to the remaining part as discussed in Remark 1.

The proposed algorithm is based on the fact in Graves, Hooker and Ramsay (2009) that, given $\phi_l(t)$, $1 \le l < k$, and r_{ik}'s, the kth component function $\phi_k(t)$ is the minimizer of the objection function

$$E\big\|Y_i(t) - r_{ik}\phi_k(t)\big\|^2, \tag{3}$$

subject to the constraints that $\|\phi_k\|^2 = 1$ and $\phi_k \perp \phi_l, \forall 1 \le l < k$. And given $\phi_k(t)$, the component score is

$$r_{ik} = \langle Y_i, \phi_k\rangle = \arg\min_r \big\|Y_i(t) - r\phi_k(t)\big\|^2. \tag{4}$$

These optimizations provide a theoretical basis for estimating $\phi_k(t)$ and r_{ik} iteratively and sequentially.

Naturally, a sample version of the objective function (3) can be constructed by

$$\frac{1}{\sum_{i=1}^N m_i}\sum_{i=1}^N\sum_{j=1}^{m_i}\big|Y_{ij} - r_{ik}\phi_k(T_{ij})\big|^2.$$

Moreover, to estimate $\phi_k(t)$, we approximate it through B-spline approximations, that is, there exists a $\boldsymbol{\alpha}_k \in \mathbb{R}^{\ell_N}$, such that $\phi_k(t) \approx \boldsymbol{\pi}(t)^T\boldsymbol{\alpha}_k$, where $\boldsymbol{\pi}(t) = \{\pi_1(t), \ldots, \pi_{\ell_N}(t)\}^T$ are ℓ_N B-spline basis functions given the specific knots and order. de Boor (1978) showed that any smooth function can always be well approximated by a B-spline representation with a sufficient number of knots. The selection of knots and order in practice is discussed in Remark 5. With the above approximations, we have the following working objective function:

$$\begin{aligned} D_{L^2}(\boldsymbol{\alpha}_k, R_k) &= \frac{1}{\sum_{i=1}^N m_i}\sum_{i=1}^N\sum_{j=1}^{m_i}\big|Y_{ij} - r_{ik}\boldsymbol{\pi}(T_{ij})^T\boldsymbol{\alpha}_k\big|^2, \\ &\text{s.t.} \quad \|\boldsymbol{\pi}(t)^T\boldsymbol{\alpha}_k\|^2 = 1 \text{ and } \boldsymbol{\pi}(t)^T\boldsymbol{\alpha}_k \perp \boldsymbol{\pi}(t)^T\boldsymbol{\alpha}_l, \forall 1 \le l < k, \end{aligned} \tag{5}$$

where $R_k = (r_{1k}, \ldots, r_{Nk})^T$ is the vector of the kth component scores.

In what follows, we present a sequential and iterative algorithm to estimate $\boldsymbol{\alpha}_k$ and R_k in (5). Our proposed alogrithm is inspired by the iterative least square method in Wold (1966), which was used to conduct multivariate PCA. A similar algorithm in alignment with robust regressions was studied in Chen, He and Wei (2008). However, our algorithm is the first attempt to implement such an iterative algorithm in PCA for sparse functional data.

Estimating the 1*st component*. The algorithm starts with estimating the 1st component $(\boldsymbol{\alpha}_1, R_1)$. We use $\boldsymbol{\alpha}_1^{(\nu)}$ and $R_1^{(\nu)}$ for the estimates of $\boldsymbol{\alpha}_1$ and R_1 at the νth iteration. The algorithm includes the following steps:

Step 1: *Initial values*. Generate R_1 with each of its elements following uniform $(0, 1)$ distribution and denote it as $R_1^{(0)}$.

Step 2: *Alternating regressions*. Continue from the νth iteration step with $R_1^{(\nu)}$. We obtain $\boldsymbol{\alpha}_1^{(\nu+1)}$ by

$$\boldsymbol{\alpha}_1^{(\nu+1)} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^{\ell_N}} \frac{1}{\sum_{i=1}^N m_i} \sum_{i=1}^N \sum_{j=1}^{m_i} |Y_{ij} - r_{i1}^{(\nu)} \boldsymbol{\pi}(T_{ij})^T \boldsymbol{\alpha}|^2, \tag{6}$$

and then standardize $\boldsymbol{\alpha}_1^{(\nu+1)}$ by $\frac{\boldsymbol{\alpha}_1^{(\nu+1)}}{\sqrt{\|\boldsymbol{\pi}(t)^T \boldsymbol{\alpha}_1^{(\nu+1)}\|^2}}$. The resulting $\boldsymbol{\alpha}_1^{(\nu+1)}$ satisfies $\|\boldsymbol{\pi}(t)^T \boldsymbol{\alpha}_1^{(\nu+1)}\|^2 = 1$. Next we update the component scores $R_1^{(\nu+1)}$ by

$$r_{i1}^{(\nu+1)} = \arg\min_{r\in\mathbb{R}} \sum_{j=1}^{m_i} |Y_{ij} - r\boldsymbol{\pi}(T_{ij})^T \boldsymbol{\alpha}_1^{(\nu+1)}|^2, \qquad i = 1, 2, \ldots, N. \tag{7}$$

Here (7) involves N separate regressions. Continue iterations until the following two conditions are satisfied:

1. The differences of $R_1^{(\nu)}$ and $R_1^{(\nu+1)}$, $\boldsymbol{\alpha}_1^{(\nu)}$ and $\boldsymbol{\alpha}_1^{(\nu+1)}$ are less than some small value δ_1 for all their elements;
2. The change in the objective function $D_{L^2}(\boldsymbol{\alpha}_1, R_1)$ between two consecutive iterations does not exceed a small value δ_2.

Step 3: *Solutions*. We denote the resulting estimates from step 2 as the $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{R}_1$, which are the estimates for $\boldsymbol{\alpha}_1$ and R_1.

It is easy to see that the objective function $D_{L^2}(\boldsymbol{\alpha}_1, R_1)$ is monotonically nonincreasing at each iterative step, and the algorithm will converge to a local minimizer.

Estimating the k*th component with* $k > 1$. When we move to the kth component $(\boldsymbol{\alpha}_k, R_k)$ with $k > 1$, we need to solve the constrained objective function (5). A numerical algorithm directly incorporating such constraints is not straightforward. However, if subtracting $\sum_{l=1}^{k-1} \widehat{r}_{il} \boldsymbol{\pi}(T_{ij})^T \widehat{\boldsymbol{\alpha}}_l$ from Y_{ij}, and denoting the resulting residuals as $\xi_{ij}^{(k-1)}$, we then have the following alternative but equivalent objective function:

$$\frac{1}{\sum_{i=1}^N m_i} \sum_{i=1}^N \sum_{j=1}^{m_i} |\xi_{ij}^{(k-1)} - r_{ik} \boldsymbol{\pi}(T_{ij})^T \boldsymbol{\alpha}_k|^2, \tag{8}$$

subject to the only constraint $\|\boldsymbol{\pi}(t)^T \boldsymbol{\alpha}_k\| = 1$. The equivalence between (8) and (5) comes from the fact that the component function $\phi_k(t)$ is also the minimizer of

$E\|Y_i^{(k-1)}(t) - \langle Y_i^{(k-1)}, \phi_k\rangle\phi_k(t)\|^2$, where $Y_i^{(k-1)}(t)$ is $Y_i(t) - \sum_{l=1}^{k-1}\langle Y_i, \phi_l\rangle\phi_l(t)$. The new objective function (8) of $(\boldsymbol{\alpha}_k, R_k)$ is the same in format as the one for $(\boldsymbol{\alpha}_1, R_1)$. Therefore, estimating $(\boldsymbol{\alpha}_k, R_k)$ can be achieved in a similar fashion as $(\boldsymbol{\alpha}_1, R_1)$. The only difference is at each iteration step, we need to orthogonalize $\boldsymbol{\pi}^T(t)\alpha_k$ against the previously estimated $\boldsymbol{\pi}^T(t)\widehat{\alpha}_l, \forall l < k$ to further improve the computational stability. The numerical details of orthogonalization are provided in Remark 4. When the observations of the growth paths are sufficiently dense, the orthogonality holds automatically without the orthogonalization step. The convergence and nonincreasing property also hold for each k. The R program for the proposed algorithm is provided in the supplemental documents Zhang and Wei (2015).

At last, to determine the number of necessary components K, we propose a model adequacy measure that is an analog of R^2 from Croux et al. (2003). It measures the total variability explained by the first K components, that is,

$$R^2(K) = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1}^{m_i}\{Y_{ij} - \sum_{k=1}^{K}\widehat{r}_{ik}\boldsymbol{\pi}(T_{ij})^T\widehat{\boldsymbol{\alpha}}_k\}^2}{\sum_{i=1}^{N}\sum_{j=1}^{m_i}Y_{ij}^2}. \tag{9}$$

We stop the estimation algorithm when $R^2(K)$ is sufficiently large. The PCA approximation of $Y_i(t)$ can be returned as $\widehat{Y_i(t)} = \sum_{k=1}^{K}\widehat{r}_{ik}\boldsymbol{\pi}(t)^T\widehat{\boldsymbol{\alpha}}_k$.

REMARK 1. The above estimation algorithm assumes that $U(t) = 0$, hence, one needs to properly center the growth paths $Y_i(t)$'s before using the algorithm. We propose to estimate the mean function $U(t) \hat{=} E\{Y(t)\}$ using nonparametric methods, such as B-spline smoothing and local polynomial smoothing, which provide uniform consistent estimators of the population mean as shown in Hansen (2008), de Boor (1978) and Fan and Gijbels (1996). Therefore, the algorithm can be applied to centered data $Y_{ij}^* = Y_{ij} - \widehat{U}(T_{ij})$, where $\widehat{U}(t)$ is the estimate of $U(t)$. Here Y_{ij}^* are asymptotically equivalent to the truly centered data as proved in Han and Lim (2010).

REMARK 2. In step 2 of our proposed algorithm, we standardize $\boldsymbol{\alpha}_k^{(\nu)}$ by $\frac{\boldsymbol{\alpha}_k^{(\nu)}}{\sqrt{\|\boldsymbol{\pi}(t)^T\boldsymbol{\alpha}_k^{(\nu)}\|^2}}$ in each iteration. The standardization step is to meet the constraint that $\|\boldsymbol{\pi}(t)^T\boldsymbol{\alpha}_k\|^2 = 1$. It does not alter the value of objection function $D_{L^2}(\boldsymbol{\alpha}_k, R_k)$ since $r_{ik}\boldsymbol{\alpha}_k^T = r_{ik}cc^{-1}\boldsymbol{\alpha}_k^T$ for any nonzero real number c.

REMARK 3. The proposed algorithm can also be used to obtain singular value decomposition of functional data. Let $\mathbf{Y}(t) = \{Y_1(t), \ldots, Y_N(t)\}^T$, $\mathbf{R} = (R_1, R_2, \ldots)$, and $\Phi(t) = \{\phi_1(t), \phi_2(t), \ldots\}$, then the decomposition (2) can be written as $\mathbf{Y}(t) = \mathbf{R}\Phi(t)$. If we further decompose $\mathbf{R} = UD$, where D is a diagonal matrix, we yield the singular value decomposition for $\mathbf{Y}(t)$, that is, $\mathbf{Y}(t) = UD\Phi(t)$. This step can be easily incorporated to the algorithm, but further decompositions of $\mathbf{R}$ are out of interest in our context.

REMARK 4. Let $\mathbf{W} = \int \boldsymbol{\pi}(t)\boldsymbol{\pi}(t)^T dt$, where $\boldsymbol{\pi}(t) = \{\pi_1(t), \ldots, \pi_{\ell_N}(t)\}^T$ are the given B-spline basis functions. $\mathbf{W}$ is a $\ell_N \times \ell_N$ matrix. Each element of $\mathbf{W}$ is the inner product of two basis functions, which can be calculated from numerical integrations. Since $\mathbf{W}$ is a positive-definite matrix, it can be decomposed as the cross-product of $\mathbf{W}^{1/2}$. In this way, $\boldsymbol{\pi}(t)^T\boldsymbol{\alpha}_k \perp \boldsymbol{\pi}(t)^T\boldsymbol{\alpha}_l$ is equivalent to $(\mathbf{W}^{1/2}\boldsymbol{\alpha}_k)^T(\mathbf{W}^{1/2}\boldsymbol{\alpha}_l) = 0$. The orthogonalization of $\mathbf{W}^{1/2}\boldsymbol{\alpha}_k$ against $\{\mathbf{W}^{1/2}\boldsymbol{\alpha}_l\}_{l=1}^{k-1}$ can be achieved through Gram–Schmidt orthonormalization from Trefethen and Bau (1997), which projects $\mathbf{W}^{1/2}\boldsymbol{\alpha}_k$ into the orthogonal space spanned by $\{\mathbf{W}^{1/2}\boldsymbol{\alpha}_l\}_{l=1}^{k-1}$, obtains the projection as $\mathbf{W}^{1/2}\boldsymbol{\alpha}_k^{\text{proj}}$, and hence has $\boldsymbol{\alpha}_k^{\text{proj}}$ as the orthogonalized $\boldsymbol{\alpha}_k$. In each of the iterative steps, we implement such orthogonalization to update $\boldsymbol{\alpha}_k^{(v)}$, which makes the final solution of $\widehat{\alpha}_k$ satisfy $\boldsymbol{\pi}(t)^T\widehat{\alpha}_k \perp \boldsymbol{\pi}(t)^T\widehat{\alpha}_l, l < k$.

REMARK 5. In practice, we choose the knots of B-spline basis functions to be $q-1$ equally spaced quantiles of pooled time points, that is, $\frac{1}{q}, \frac{2}{q} \cdots \frac{q-1}{q}$th quantiles. In this way, the B-spline basis functions are determined by q and order. Since there are only two parameters, it is straightforward to choose them by 5-fold cross-validation using AIC or BIC criterion. Based on our numerical experience, the results are not sensitive to the exact locations of knots.

REMARK 6. The proposed algorithm has a lack of consistency of results for the estimated principal component functions and scores under the sparsity setting in this paper. A weak asymptotic result for the principal component functions under restrictive assumptions exists.

2.3. *The construction of growth charts for growth paths.* Through the proposed PCA algorithm, we can approximate $Y_i(t)$ as $\widehat{U}(t) + \sum_{k=1}^{K} \widehat{r}_{ik}\boldsymbol{\pi}(t)^T\widehat{\boldsymbol{\alpha}}_k$. Hence, the percentile ranks of $Y_i(t)$ can be identified by estimating the multivariate quantiles of $(\widehat{r}_{i1}, \ldots, \widehat{r}_{iK})$. Multivariate quantiles consider the joint distribution of components scores and bring additional insights in screening growth patterns. The individual percentile ranks determined by component scores enable the comparisons among subjects, which can be useful for pediatric practice. For example, subject A is at the 95th percentile and subject B is at the 97th percentile. Using the percentile ranks, a pediatrician can prioritize the work by examining the health status of subject B first, since subject B is more likely to have health issues given its higher percentile rank.

Due to the lack of natural ordering in a multidimensional space, there is no universally preferred definition of multivariate quantiles, but various ideas have been developed in the literature. For example, Liu, Parelius and Singh (1999) and Zuo and Serfling (2000) used multivariate quantile functions based on the half-space depth functions. Other approaches have been given by Parzen (1979), Abdous and Theodorescu (1992), Hettmansperger, Nyblom and Oja (1992), Chaudhuri

(1996), Koltchinskii (1997), Chakraborty (2003), McDermott and Lin (2007) and Wei (2008). Serfling (2002) presented a nice survey of multivariate quantile functions and outlined the probabilistic properties that a multivariate quantile function should have.

In our case, the joint distribution of $(\widehat{r}_{i1}, \ldots, \widehat{r}_{iK})$ is unlikely to follow a certain parametric distribution due to the complexity of sparse functional data. Therefore, we propose to determine their multivariate quantiles nonparametrically using Wei (2008), since this method is also motivated from growth chart problems, and measuring the spatial "outlyingness" of an observation relative to a center, which is the essential part of growth chart studies. Wei (2008) converts the component scores into the polar coordinate system and builds the quantile contours by nonparametrically regressing the radiuses with respect to the angles at various quantile levels. Then, by building a sequence of nested multivariate quantile contours of the K component scores, our growth chart can be constructed and used to determine the percentile ranks of growth paths.

Suppose we want to use our constructed growth chart to screen a growth path of a new subject, including m_* observed measurements, $\{T_{*j}, Y_{*j}\}_{j=1}^{m_*}$. We first obtain its component scores $\{r_{*1}, \ldots, r_{*K}\}$ by the following least square regression:

$$\min_{r_1,\ldots,r_K \in \mathbb{R}} \frac{1}{m_*} \sum_{j=1}^{m_*} \left| Y_{*j} - \widehat{U}(T_{*j}) - \sum_{k=1}^{K} r_k \widehat{\phi}_k(T_{*j}) \right|^2, \tag{10}$$

where $\widehat{U}(t)$ and $\widehat{\phi}_k(t)$ are estimated from the reference growth data. By the estimated component scores, this subject can then be located on the constructed growth chart. If it stays outside an extreme quantile contour, such as the 0.95th quantile, we say that its growth path is more unusual than at least 95% of its peers, hence it can be singled out for further clinical investigations.

2.4. *Incorporating covariate effects.* Since incorporating subject level information, such as parental information and ethnicity, can enhance screening performance, we extend our proposed method to include a covariate X. Suppose the reference growth data consist of $\{(Y_{ij}, T_{ij}, X_i), i = 1, \ldots, N, j = 1, \ldots, m_i\}$, where X_i is the covariate of the ith subject. We assume that the measurement Y_{ij} is observed from

$$Y_{ij} = Y_i(T_{ij}, X_i) + \varepsilon_{ij},$$

where $Y_i(t, x)$ is the underlying growth path for the ith subject, and depends on both age t and covariate x. By extending the Karhunen–Loeve decomposition, we can write

$$Y_i(t, x) = U(t, x) + \sum_{k=1}^{\infty} r_{ik} \phi_k(t, x), \qquad t \in \mathcal{T}, \tag{11}$$

where $U(t,x)$ is the mean function, $\phi_k(t,x)$'s are pair-wise orthogonal component functions, and r_{ik}'s are individual component scores with respect to $\phi_k(t,x)$. Following similar ideas in Section 2.2, we extend the working objective function (5) as follows:

$$D_x(r_{ik},\boldsymbol{\alpha}_k) \hat{=} \frac{1}{\sum_{i=1}^N m_i}\sum_{i=1}^N\sum_{j=1}^{m_i}\left|Y_{ij}-r_{ik}\boldsymbol{\pi}(T_{ij})^T\boldsymbol{\alpha}_k\boldsymbol{\mu}(X_i)\right|^2 \quad \text{s.t.,} \tag{12}$$

$$\int\left\{\boldsymbol{\pi}(t)^T\boldsymbol{\alpha}_k\boldsymbol{\mu}(x)\right\}^2 dt = 1; \tag{13}$$

$$\int\left\{\boldsymbol{\pi}(t)^T\boldsymbol{\alpha}_k\boldsymbol{\pi}(x)\right\}\left\{\boldsymbol{\pi}^T(t)\boldsymbol{\alpha}_l\boldsymbol{\mu}(x)\right\}dt = 0 \qquad \forall 1\le l<k. \tag{14}$$

Here $\boldsymbol{\pi}(t)^T\boldsymbol{\alpha}_k\boldsymbol{\mu}(x)$ provides the approximation of $\phi_k(t,x)$, where $\boldsymbol{\pi}(t)$ is the B-spline basis functions for t as in Section 2.2, $\boldsymbol{\mu}(x)=\{\mu_1(x),\ldots,\mu_{\ell_x}(x)\}^T$ is a set of covariate functions, and $\boldsymbol{\alpha}_k$ becomes a $\ell_N\times\ell_x$ matrix instead of a vector. The simplest choice of covariate functions $\boldsymbol{\mu}(x)$ is $(1,x)^T$, which implicitly assumes the component functions are linear in x for any given t. If the linearity assumption does not hold, one could consider including quadratic terms of x or even choosing $\boldsymbol{\mu}(x)$ as B-spline basis functions to avoid any parametric assumption. Since $\boldsymbol{\pi}(t)^T\boldsymbol{\alpha}_k\boldsymbol{\mu}(x)$ is still a linear function of $\boldsymbol{\alpha}_k$, we can implement the similar iterative algorithm in Section 2.2 by alternatively updating $\boldsymbol{\alpha}_k$ and r_{ik}. The major differences in each iteration come from the standardization and orthogonalization of $\boldsymbol{\pi}(t)^T\boldsymbol{\alpha}_k\boldsymbol{\mu}(x)$ in order to meet constraints (13) and (14), details of which are provided in Zhang (2012). Similarly, the covariate adjusted algorithm is conducted sequentially, and stopped when reaching an appropriate number of components K, which is determined by the extended R^2, that is, $1-\frac{\sum_{i=1}^N\sum_{j=1}^{m_i}\{Y_{ij}-\sum_{k=1}^K\widehat{r}_{ik}\boldsymbol{\pi}(T_{ij})^T\widehat{\boldsymbol{\alpha}}_k\boldsymbol{\mu}(X_i)\}^2}{\sum_{i=1}^N\sum_{j=1}^{m_i}(Y_{ij})^2}$. Then the underlying growth path $Y_i(t,X_i)$ can be well approximated by the first several component functions, and hence determined by its component scores. Therefore, the growth chart for screening growth path can be constructed and implemented in a similar fashion as the one described in Section 2.3.

3. Application examples.

3.1. *Growth charts for screening pubertal growth paths.* In this section we illustrate our proposed screening method using part of a Finnish national growth data set from Pere (2000). The data consist of longitudinal height measures of 553 girls (ages 9–16) and 518 boys (ages 11–19) during puberty, as shown in Figure 2. The median number of measurements for each subject is 6. The analysis is stratified by gender. We apply the proposed regression based PCA using quadratic B-splines with internal knots 11 and 13.56. The resulting first two component functions are plotted in Figure 3 for girls and Figure 4 for boys. In both

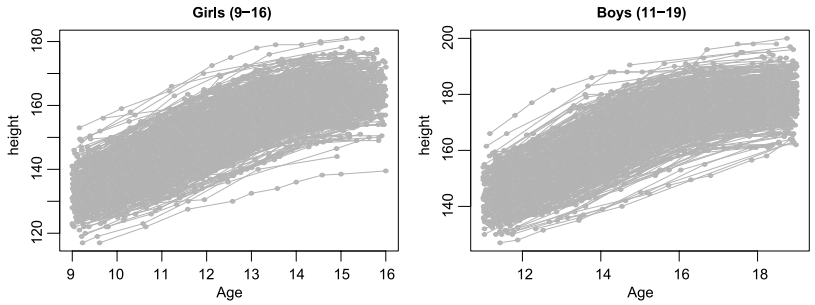


FIG. 2. *Part of a Finnish national growth data from Pere (2000). The data include the longitudinal height measurements for* 553 *girls* (*left*) *from ages* 9 *to* 16 *and* 518 *boys* (*right*) *from ages* 11 *to* 19. *The* y*-axis is height and the* x*-axis is age. The dots are the observed height measurements.*

cases, they count for 90% variability of the growth paths based on the proposed R^2 measure (9).

In both genders, we find that the first component function $\phi_1(t)$ reflects the overall growth scale, while the second one $\phi_2(t)$ coincides well with the puberty growth velocity pattern. The second component function increases rapidly starting around age 11 and stabilizes after age 15 for girls [Figure 3(b)], while a similar patten is found between age 14 and age 18 for boys [Figure 3(b)]. This difference in $\phi_2(t)$ is biologically reasonable since the puberty of boys begins later than girls. Therefore, the corresponding principal component scores have a nice biological interpretation. A subject with a higher r_{i1} tends to be taller than most of his or her peers, while a subject with a higher r_{i2} may experience rapid pubertal growth. The growth charts are constructed based on the first two component scores, as shown in Figure 5(a) for girls and Figure 5(b) for boys. Such charts provide a

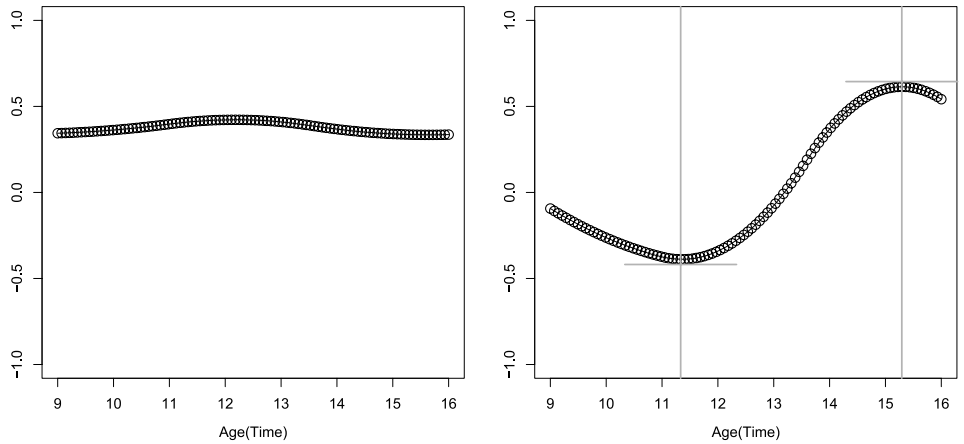


FIG. 3. *The estimated first two component functions* $\widehat{\phi}_1(t)$ (a) *and* $\widehat{\phi}_2(t)$ (b) *for girls.*

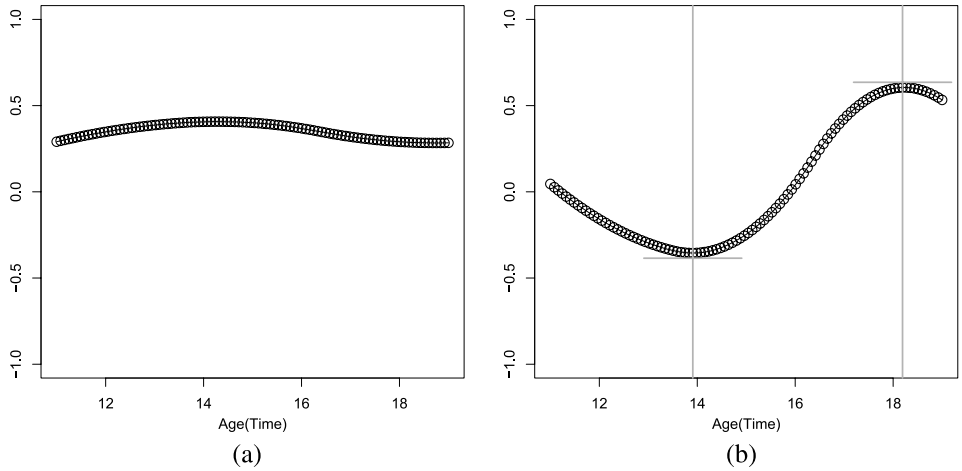


FIG. 4. *The estimated first two component functions* $\widehat{\phi}_1(t)$ (a) *and* $\widehat{\phi}_2(t)$ (b) *for boys.*

convenient visual tool for screening potentially unusual growth patterns. In both figures, the x axis represents the first component score and the y axis represents the second ones. Bivariate quantile contours at quantile levels 0.5, 0.75 and 0.95 are added to determine the individual percentile ranks. The individuals staying outside the 0.95th quantile contour have more outlying component scores than at least 95% of their peers. Hence, they will be screened out for further clinical investigations.

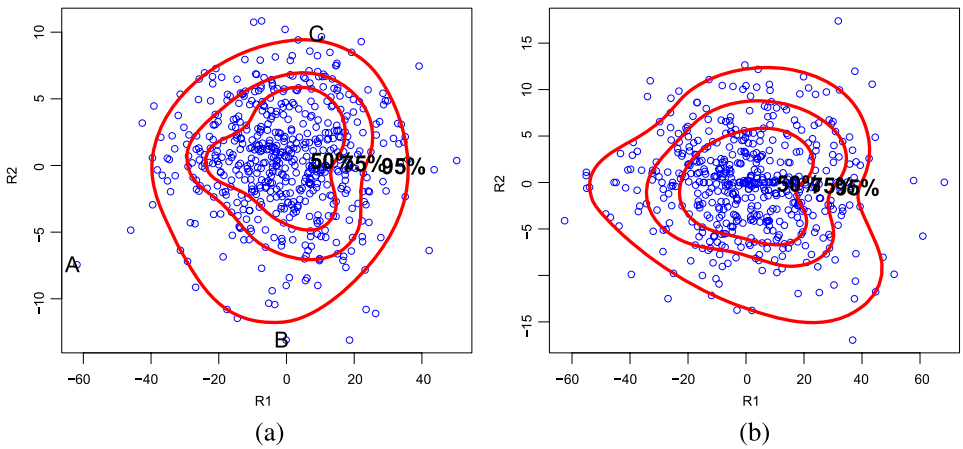


FIG. 5. *The bivariate plot of the first two component scores for girls* (a) *and boys* (b). *The x axis represents the first component score and the y axis represents the second component score. The contours from inside to outside are the bivariate quantile contours at quantile levels* 0.5, 0.75 *and* 0.95. *The points labeled "*A*" and "*B*" in* (a) *are two selected girls whose first two component scores fall outside the* 0.95*th quantile contour.* (a) *The growth chart for girls.* (b) *The growth chart for boys.*

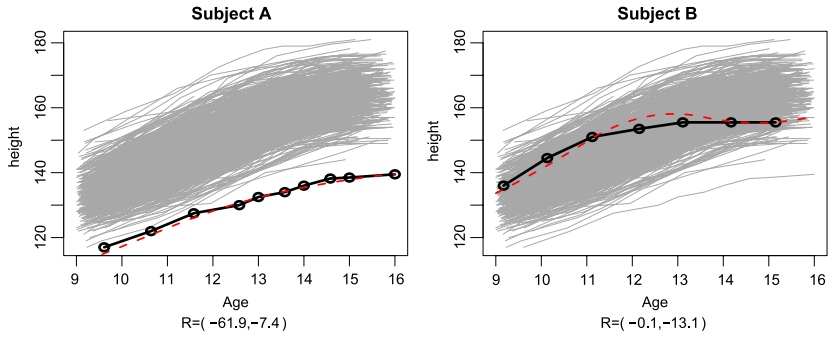


FIG. 6. *The observed growth paths of two extreme girls, girl* A (a) *and girl* B (b) *in Figure* 5(a). *The black dots are the original height measurements, and the dashed lines are the estimated growth paths. The gray background curves are all the growth paths from the Finnish growth data for girls.*

To illustrate the screening performance of our constructed growth charts, we select two girls, A and B, who are outside the 0.95th quantile contour in Figure 5(a), and further examine their growth paths as shown in Figure 6. In Figure 6, the black dots are the original height measurements, and the dashed lines are the estimated underlying growth path $Y_i(t)$. The gray curves in the background are all the growth paths from the data. According to Figure 5(a), girl A has small component scores in both directions, while girl B has an average first component score, but a very low second component score. Consequently, as shown in Figure 5, girl A is shorter and slower than most of her peers; girl B has normative height, but apparently fails to gain enough height during her puberty. In both cases, the unusual growth patterns detected by our proposed growth charts are confirmed by empirical observations of the growth paths.

Comparison to existing growth charts. As we illustrate in the Introduction, screening entire growth paths may bring new insights in monitoring human growth. The outlying girl C in Figure 5(a) is one example. Figure 7 provides the observed growth path of girl C. Her height starts around the median at the age of 9 and gradually increases to the upper percentile by the age of 16.

We first screen each of her measurements (black dots) using conventional growth charts and conditional growth charts. Specifically, following conventional growth charts from Wei et al. (2006), we estimate the 0.025th and 0.975th percentiles that are conditioned only on her ages (squares in Figure 7). And following conditional growth charts from Wei et al. (2006), we estimate the same reference percentiles conditioned on both her ages and prior measurements (triangles in Figure 7).

As shown in Figure 7, all of her height measurements are within the normal ranges of both conventional and conditional growth charts. Therefore, when these

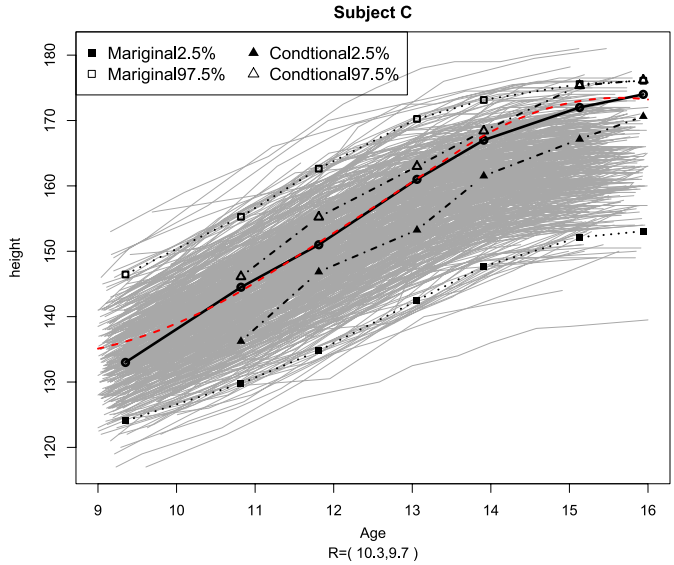
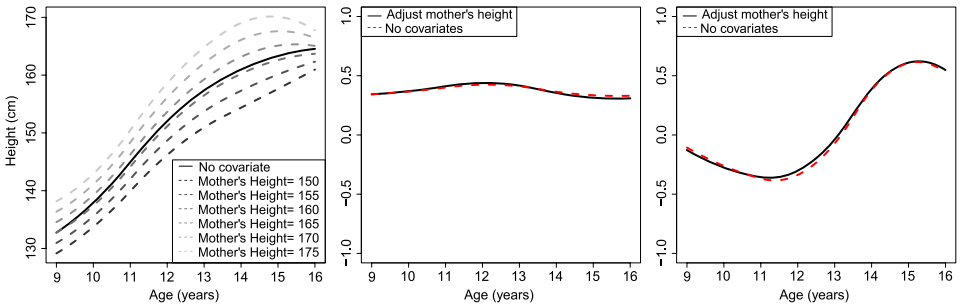


FIG. 7. *The observed growth path of one extreme girl* (*girl* C) *in Figure* 5(a). *The black dots are the original height measurements and the dashed line is the estimated growth path. The gray background curves are all the growth paths from the Finnish growth data for girls. The squares are the estimated* 0.975*th* (*open squares*) *and* 0.025*th* (*solid squares*) *quantiles from the unconditional growth chart. The triangles are the estimated* 0.975*th* (*open squares*) *and* 0.025*th* (*solid squares*) *quantiles from the conditional growth chart.*

two growth charts are used to screen her height one at a time, each of her height measurements is considered as normative. However, when we screen her entire growth path using the proposed method, girl C is screened out by the 0.95th quantile contour since her second component score appears unusually large. It is consistent with the fact that she has been growing fast consecutively over her entire puberty. This example shows that the proposed method provides informative insights on growth pattern by considering entire paths.

3.2. *Growth charts conditioned on mother's height.* Parental heights usually have strong associations with their children's growth. In this section we incorporate mother's height into the model and examine the pubertal growth of the Finnish teenage girls. The data set used here is a subset of girls' data in Section 3.1, including 444 girls with mother's height information available and at least 5 measurements between ages 9 and 16. To make the comparisons, we apply our proposed method, both with covariate and without covariate, to the data. We choose $\boldsymbol{\mu}(x)$ to be $\boldsymbol{\mu}(x) = (1, x)^T$. Under this parameterization, $U(t, x) = U_1(t) + xU_2(t)$, $\phi_1(t, x) = \phi_{11}(t) + x\phi_{12}(t)$, and $\phi_2(t, x) = \phi_{21}(t) + x\phi_{22}(t)$. The unknown functions $U_1(t)$, $U_2(t)$, $\phi_{11}(t)$, $\phi_{12}(t)$, $\phi_{21}(t)$ and $\phi_{22}(t)$ are all approximated using quadric B-splines with internal knots equal to $1/3$ and $2/3$ quantiles of pooled times.



(a) (b) (c)

FIG. 8. (a) *The estimated mean functions from the covariate adjusted model* (*dashed lines*) *and the model without covariate* (*solid line*). *The dashed lines are the estimated mean functions conditioned on six different mother's heights. The lines from the lightest gray to the darkest gray represent* 150 *cm*, 155 *cm*, 160 *cm*, 165 *cm*, 170 *cm and* 175 *cm*, *respectively.* (b), (c) *The estimated first two component functions from the covariate adjusted model* (*dashed lines*) *and the model without covariate* (*solid lines*). (a) *Estimated location functions.* (b) $\widehat{\phi}_1(t)$ *for girls.* (c) $\widehat{\phi}_2(t)$ *for girls.*

We use a bootstrap to test whether the covariate associated functions $U_2(t)$, $\phi_{12}(t)$ and $\phi_{22}(t)$ are equal to zero at any t, which is essentially testing whether the corresponding B-spline coefficients are equal to 0. More details can be found in Zhang (2012). The resulting p-values indicate that the mother's height is significantly related to $U(t,x)$ (p-value ≤ 0.0001), while $\phi_1(t,x)$ and $\phi_2(t,x)$ are insignificant (p-values equal to 0.72 and 0.59). We hence simplify the covariate adjusted model to

$$Y_i(t, X_i) \approx U_1(t) + xU_2(t) + r_{i1}\phi_{11}(t) + r_{i2}\phi_{21}(t).$$

In Figure 8(a), the solid line is the estimated mean function without considering mother's height, and the dash lines are the expected growth paths conditioned on six different mother's heights which are 150 cm, 155 cm, 160 cm, 165 cm, 170 cm and 175 cm (from darkest gray to the lightest grey), respectively. Covariate adjusted mean functions show that with the increase of mother's height, the expected body sizes and growth rates both tend to increase as well. We also observe the expected growth path conditioned on 160 cm is close to the expected growth path of the whole population. The explanation is that the average of mother's height in this data set is 161.6 cm, which is close to 160 cm. As shown in Figures 8(b)–(c), the estimated component functions from both models are very close to each other. However, due to the difference in the mean functions, the distributions of individual component scores are fairly different between the two models. Figure 9 plots the bivariate quantile contours estimated from two sets of component scores. We say that Figure 9(a) is the covariate adjusted growth chart for puberty growth paths and Figure 9(b) is the marginal one.

Two girls, D and E, are selected from the sample and placed against the two growth charts. The growth path of girl D is considered as unusual in the marginal

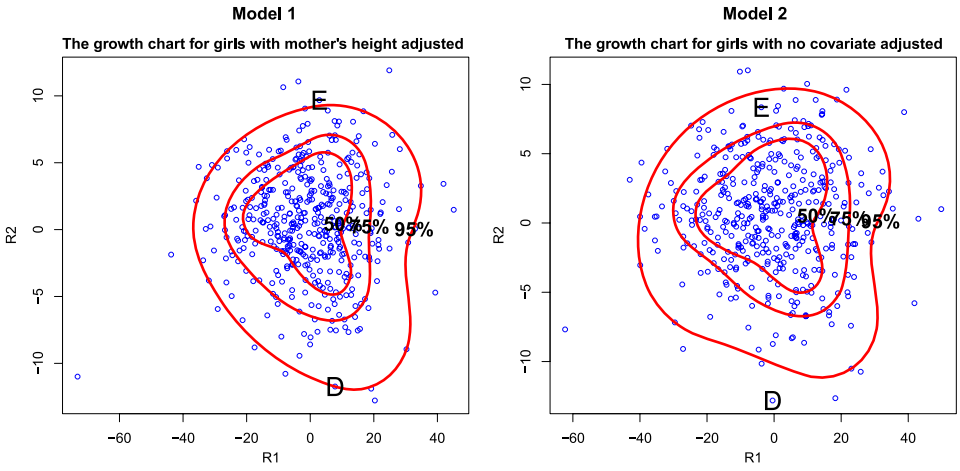


FIG. 9. *The bivariate plots of the first two component scores for the covariate adjusted model* (a) *and the model without covariate* (b). *The x axis represents the first component score and the y axis represents the second component score. The contours from inside to outside are the bivariate quantile contours at quantile levels* 0.5, 0.75 *and* 0.95.

growth chart, but not in the covariate adjusted one. In contrast, the growth path of girl E is only considered as unusual in the covariate adjusted growth chart, but not in the marginal one. Figures 10 and 11 provide their growth paths (black solid lines and dots) for further investigations. In Figure 10(a), we compare the target

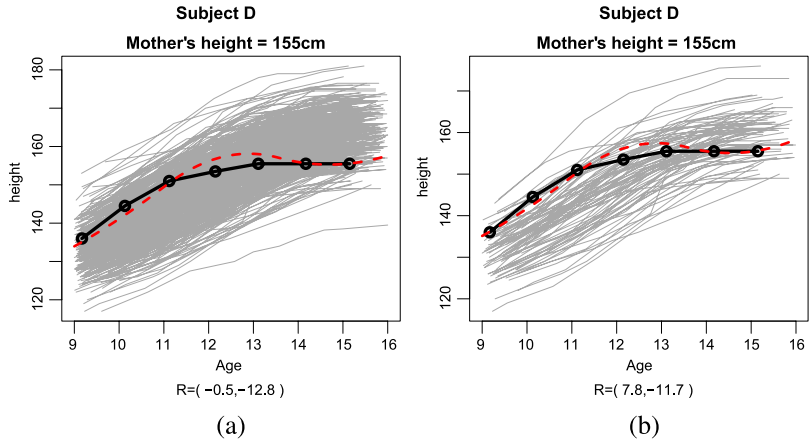


FIG. 10. *The observed growth path of girl* D *in bivariate plots Figure 9. The black dots are the original height measurements. The gray background curves in* (a) *are all the growth paths from this data set. The gray background curves in* (a) *are the growth paths of the individuals with mother's height from* 153 *cm to* 155 *cm.*

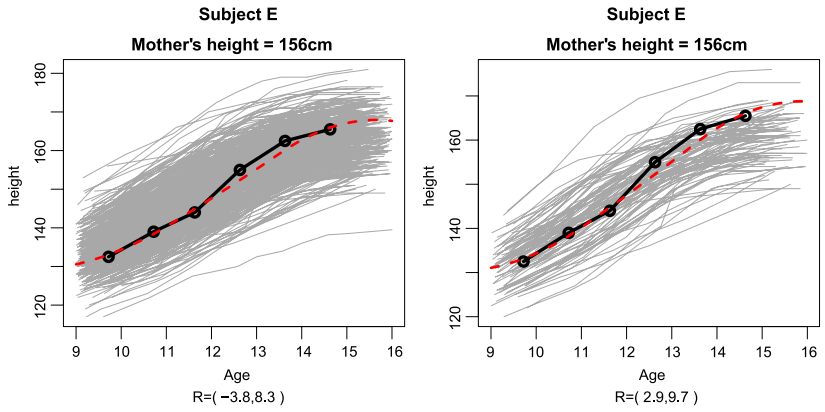


FIG. 11. *The observed growth path of girl* E *in bivariate plots Figure 9. The black dots are the original height measurements. The gray background curves in* (a) *are all the growth paths from this data set. The gray background curves in* (a) *are the growth paths of the individuals with mother's height from* 154 *cm to* 158 *cm.*

paths to all the growth paths in the sample (gray curves), while in Figures 10(b), we compare them only to those (gray curves) who have similar mother's heights ($\pm$2cm). We find that girl D has grown unusually slow from ages 12 to 16 compared to others in the entire sample. That explains why girl D has an unusually low second component score in the marginal growth chart. However, if one restricts to those whose mothers have heights around 155 cm, her slow puberty growth is less extreme, as we observe more similar slow growth patterns in this subset. Subject E has normative body sizes and growth rates according to the marginal growth chart, but has excessive growth based on the covariate adjusted chart. Examining her growth path in Figure 11, we find that she has consecutive years of fast growth from ages 12 to 15. This fast growth appears to be more extreme when being compared to those whose mothers have similar heights. In this case, we would have missed the excessive growth of girl E if we did not take her mother's height into consideration. These examples show that incorporating subject level information, especially parental information, might lead to improvements in screening growth paths.

4. Numerical investigations.

4.1. *Finite sample performance.* In this section we present a numerical simulation study to illustrate the finite sample performance of the proposed PCA method in comparison to the alternative Yao, Müller and Wang (2005) and MLE methods. For MLE methods, we use the fpca R package based on Peng and Paul (2009) since it provided an improved fitting of James, Hastie and Sugar (2000).

We consider the following model to generate the simulation data:

$$Y_{ij} = Y_i(T_{ij}) = U(T_{ij}) + r_{i1}\phi_1(T_{ij}) + r_{i2}\phi_2(T_{ij}) + \varepsilon_{ij},$$

where $\phi_1(t)$, $\phi_2(t)$ and $U(t)$ are chosen to be the estimated functions for girls in Section 3.1. We consider the following two distributions for (r_{i1}, r_{i2}). In setting 1, we generate them from the empirical distribution of the estimated first two component scores for girls in Section 3.1. In setting 2, we generate them from a bivariate normal distribution with sample means and covariance estimated from the first two component scores for girls in Section 3.1. Both settings try to mimic growth paths of the Finnish data for girls, while a more restrictive parametric assumption is made in setting 2. For each of the above two settings, we generate 20 Monte Carlo samples. Each sample includes $N = 500$ random curves. Each one consists of $m_i = 6$ observations with the observed time T_{ij} uniformly distributed on [9, 16].

For each sample, we use the proposed method, Yao, Müller and Wang (2005), and the MLE method to conduct PCA. We first estimate $U(t)$ using nonparametric regression and then apply the three methods to the centered data $Y_{ij}^* = Y_{ij} - \widehat{U}(T_{ij})$ to estimate component functions. The selection of tuning parameters for all three algorithms is described as the following. Because both our method and the MLE method from Peng and Paul (2011) use B-spline functions to represent component functions, we choose the same set of basis functions for both methods, that is, the quadratic B-spline basis functions with the 1/3th and 2/3th quantiles of the pooled times as the internal knots. Yao, Müller and Wang (2005) relied on estimating the variance and covariance by two-dimensional local polynomial smoothing. Its smoothing parameters are determined by minimizing the AIC type criterion, that is, $N \times \log\{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{m_i}\sum_{j=1}^{m_i}(Y_{ij} - \widehat{Y}_{ij})^2\} + 2p$, where p is the number of parameters and $\widehat{Y}_{ij}$ is the predicted Y_{ij}. All codes for the simulations are written in R language and run under R version 3.0.0 on a machine with Intel(R) Xeon(R), CPU 3.20 GHz and 16 GB RAM. On average, the running time to conduct PCA for one Monte Carlo sample is 17 seconds for our proposed method, 18 seconds for the MLE method, and 30 seconds for Yao, Müller and Wang (2005).

To evaluate the estimation performance of the three methods, we calculate relative integrate squares errors (RISE) for both $\phi_1(t)$ and $\phi_2(t)$, where RISE for estimating a target function $g(t)$ is defined as $\frac{\|g(t)-\widehat{g}(t)\|^2}{\|g(t)\|^2}$, and $\widehat{g}(t)$ is the estimate. RISE can be considered as noise to signal measurements. The integrations in RISE are evaluated using the left Riemann sum [Thomas, Finney and Weir (1988)] with the equal partition of the whole interval into 100 small intervals. Table 1 provides the summary of RISEs under both settings. As shown in Table 1, all three methods perform well in estimating component functions, although Yao, Müller and Wang (2005) have slightly larger means and standard deviations.

We further evaluate the estimation errors of component scores r_{ik} among the three methods. For each Monte Carlo sample, we calculate relative mean square error (RMSE), defined as $\frac{\sum_{i=1}^{N}(r_{ik}-\widehat{r}_{ik})^2/N}{s^2(r_{ik})}$, where $\widehat{r}_{ik}$ is the estimator of r_{ik} and

TABLE 1
The summary of RISEs for the three sparse functional PCA methods

	Means (standard deviations) of RISE		
	Yao et al. (2005)	**The MLE method**	**The proposed method**
Setting 1: $(r_{i1}, r_{i2}) \sim$ Empirical distribution			
RISE of $\phi_1(t)$	0.0061 (0.0017)	0.0003 (0.0003)	0.0004 (0.0005)
RISE of $\phi_2(t)$	0.0955 (0.0545)	0.0022 (0.0009)	0.0020 (0.0015)
Setting 2: $(r_{i1}, r_{i2}) \sim$ Bivariate normal distribution			
RISE of $\phi_1(t)$	0.0052 (0.0018)	0.0003 (0.0003)	0.0004 (0.0003)
RISE of $\phi_2(t)$	0.1076 (0.0872)	0.0023 (0.0012)	0.0027 (0.0014)

$s^2(r_{ik})$ is the sample variance of r_{ik}. RMSE measures the fraction of variance unexplained caused by estimation errors. Yao, Müller and Wang (2005) involve the estimation of the individual covariance matrix and its inverse, which can be singular or close to singular. When it happens, it can deviate the estimation of component scores r_{ik}. To make a fair comparison, we exclude the top 5% extreme square errors in the calculation of RMSE for Yao, Müller and Wang (2005). RMSEs under both settings are summarized in Table 2. All three methods work well for the 1st component with average RMSEs less than 5%. For the 2nd component scores, the average RMSEs for both our proposed method and the MLE method increase but still less than 20%, while the RMSEs for Yao, Müller and Wang (2005) tend to be slightly larger.

TABLE 2
The summary of relative mean square errors (RMSE) $\frac{\sum_{i=1}^{N}(r_{ik}-\widehat{r}_{ik})^2/N}{s^2(r_{ik})}$ *for the three sparse functional PCA methods*

	Means (standard deviations) of RMSE		
	Yao et al. (2005)*	**The MLE method**	**The proposed method**
Setting 1: $(r_{i1}, r_{i2}) \sim$ Empirical distribution			
RMSE of r_{i1}	0.05 (0.04)	0.01 (0.01)	0.02 (0.01)
RMSE of r_{i2}	0.69 (0.47)	0.13 (0.02)	0.17 (0.03)
Setting 2: $(r_{i1}, r_{i2}) \sim$ Bivariate normal distribution			
RMSE of r_{i1}	0.07 (0.05)	0.01 (0.01)	0.02 (0.01)
RMSE of r_{i2}	0.87 (0.63)	0.14 (0.03)	0.17 (0.04)

*Note: Yao et al. (2005) involve the estimation of the individual covariance matrix and its inverse, which can be singular or close to singular. When it happens, it can deviate the estimation of component scores r_{ik}. To make a fair comparison, we exclude the top 5% extreme square errors in the calculation of RMSE for Yao et al. (2005).

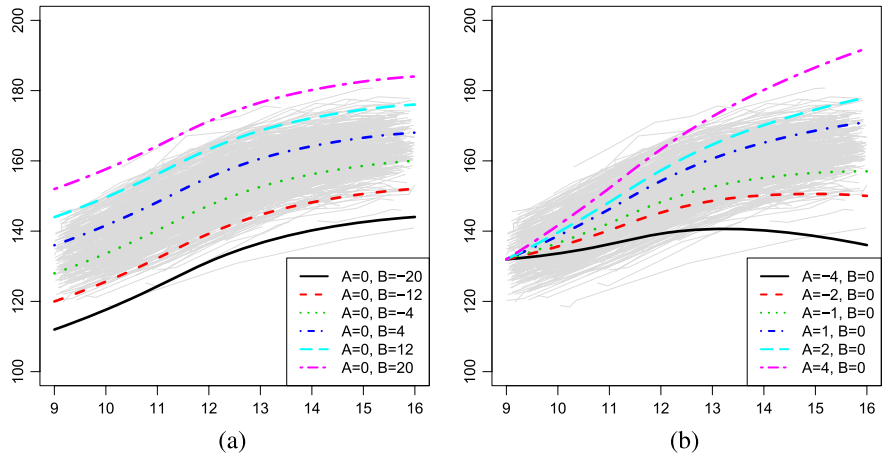


FIG. 12. (a), (b) *The selected outlying curves under different combinations of A and B. The background gray curves are simulated curves from one Monte Carlo sample.*

4.2. *Screening power.* To illustrate how sensitive the proposed method is in identifying outlying growth paths compared to the conventional and conditional growth charts, we simulate Monte Carlo samples from setting 1 as the reference growth data and build the three types of growth charts accordingly. We then simulate outlying growth paths $Z_i(t)$ from $Z_i(t) = Y_i(t) + A(t-9) + B$. Here $Y_i(t)$ follow the correct model from setting 1, and $A(t-9)+B$ is a linear contaminated term, where A provides the slope deviation and B represents the location shift. We choose A from $(-4, -2, -1, 0, 1, 2, 4)$ and B from $(-20, -12, -4, 0, 4, 12, 20)$. For each (A, B) combination, we generate 100 curves with 6 observations $Z_{ij} = Z_i(T_{ij}) + \varepsilon_{ij}$ each. Figure 12 shows the selected outlying curves (dashed lines) under several combinations of A and B. The background gray curves are from one simulated sample. The simulated curves become more outlying with the increase of either $|A|$ or $|B|$. Following the procedure in Section 2.3, we locate the simulated outlying curves in the growth charts and screen out those outside the 95th percentile contours. We also screen each of the measurements from the simulated outlying curves using the conventional and conditional growth charts. Specifically, following the conventional growth chart from Wei et al. (2006), we estimate the 2.5th and 97.5th percentiles that are conditioned only on ages. And following the conditional growth chart from Wei et al. (2006), we estimate the same reference percentiles conditioned on both the ages and prior measurements. Using the conventional and conditional growth charts, we screen out the curves with more than one measurement outside the range between the corresponding 2.5th and 97.5th percentiles. Table 3 and Table 4 summarize the percentages of curves that are screened out by the growth charts, including both means and standard deviations over 20 Monte Carlo samples. The results illustrate that all three growth charts are effective in identifying outlying growth paths when both the location

TABLE 3

The means of the percentages of outlying curves $Z_i(t)$ that are screened out by the 95th percentile contours from the proposed growth chart, the 2.5th and 97.5 percentiles from the conventional growth chart, and the 2.5th and 97.5 percentiles from the conditional growth chart for different combinations of A (slope deviation) and B (location shift)

	Means of percentages: The proposed method/The conventional growth chart/The conditional growth chart						
	$B=-20$	$B=-12$	$B=-4$	$B=0$	$B=4$	$B=12$	$B=20$
$A=-4$	100/100/100	98.8/100/100	95.9/98/99.3	93.9/94.3/98.4	91.3/89.2/96.9	86.9/77.1/92.5	87.6/83/84.9
$A=-2$	99.4/100/97.2	94.9/96.8/88.9	76.4/73.2/74.5	63.9/51.7/65.6	51.9/33.6/57.2	45/46.4/42.9	67.2/85.2/34.9
$A=-1$	98/99.2/82.7	77.8/84.2/65	40.2/39.2/45.3	24.2/20.6/36.2	18.9/15.8/28.6	33.7/50/22.1	73.8/90.3/24.4
$A=0$	86.9/95.5/64.6	46.5/62.8/44.3	11.7/15.6/28.6	5.9/9.4/12.8	10.8/19.2/22.1	47/65/24.6	86.9/95.1/31.4
$A=1$	69.5/89/64	28.1/48.2/52.1	12.2/14.2/45.5	16.7/22.8/44.5	31.4/43.2/44.7	74.4/86.1/48.4	96/98.9/53.4
$A=2$	60.6/83.7/78.8	37/42.5/76.5	39.5/34.3/75.8	52.1/55/76.4	67.2/75.2/77.9	91.6/97.2/80.7	99/99.7/82.8
$A=4$	80.3/82.5/96.1	81.3/76.8/96.4	88/90.4/97.7	91.9/95.3/98	95.2/98.3/98.1	99.1/99.8/98.6	100/100/98.7

TABLE 4

The standard deviations of the percentages of outlying curves $Z_i(t)$ that are screened out by the 95th percentile contours from the proposed growth chart, the 2.5th and 97.5 percentiles from the conventional growth chart, and the 2.5th and 97.5 percentiles from the conditional growth chart for different combinations of A (slope deviation) and B (location shift)

	Standard deviations of percentages: The proposed method/The conventional growth chart/The conditional growth chart						
	$B=-20$	$B=-12$	$B=-4$	$B=0$	$B=4$	$B=12$	$B=20$
$A=-4$	0.2/0/0	1.1/0.2/0.2	2.2/1.4/0.8	3.1/2.6/1.4	3.4/3.5/1.7	3.9/5.2/2.6	3.2/4.4/5.7
$A=-2$	0.8/0/2.4	2.5/1.8/5.1	6.1/7.6/5.8	7.6/7.3/7	7.5/5.9/7.3	7.6/5.9/7.1	5.9/4/8.3
$A=-1$	1.4/0.8/9.1	6.4/4.5/8.9	7.4/6.7/7.2	7.1/5.4/6.3	7.2/3.6/4.5	7.7/5.7/6.1	6.4/2.9/8.6
$A=0$	4.5/1.8/8.6	9.2/6/7.4	4/4.1/4.7	2.8/3.1/4.7	3.7/4.5/4.3	7.7/4.7/7.6	5.7/2.1/11.9
$A=1$	12/3/8.8	9.8/6.8/8.1	5.4/4/5.3	5.5/4.4/4.6	8.6/4.9/4.4	9.8/4/9.7	3.4/1.2/14.6
$A=2$	14.9/3.4/6.9	12.1/7/6	10.3/4.4/4.8	11.8/5.3/4.5	10.9/3.8/5.3	4.3/1.4/7.8	1.3/0.7/10.1
$A=4$	6.5/3.8/1.7	5.3/5.6/1.6	4.2/2.6/1.9	3.2/2/1.9	2/1.2/1.9	1.1/0.6/1.7	0.2/0/1.7

shift and slope deviation are very extreme ($B = -20$ and $A = -4$). The conventional growth chart is most sensitive in screening out big location shifts ($A = 0$ and $B = -20, -12, 12, 20$). The conditional growth chart works the best for detecting dramatic slope deviations ($B = 0$ and $A = -4, -2, 4, 2$). The proposed growth chart works the best for identifying the unusual growth pattern combining moderate location shift and slope deviation ($B = -4$ and $A = -2$). Among the three growth charts, the proposed method has the most reasonable type I errors (the results when A and B are both 0) with mean 5.8% (9.8% for the conventional growth chart and 12.8% for the conditional growth chart).

5. Conclusion and discussion. This paper develops a new statistical method to construct growth charts for screening entire growth paths. By considering entire growth paths, the proposed growth charts bring more informative insights into monitoring pediatric growth. When our constructed growth chart is applied to the Finnish growth data for monitoring puberty growth, it shows more effective performance in detecting possible unusual growth patterns compared to existing growth charts. Besides pediatrics, our proposed method can also be applied to other areas, such as monitoring CD4 lymphocyte counts of uninfected children born to HIV-1-infected women in HIV research, and helping determine the gene frequencies of the most common mutations in the HFE gene in genetics research.

The proposed method also contributes to the statistical methodologies. First, it provides a new way to rank longitudinal/sparse functional data. It approximates the sparse and irregularly spaced functional data through PCA and represents each individual using the resulting components scores. Then the percentile rank of each individual can be identified by applying multivariate methods to components scores. Second, the proposed regression based PCA algorithm provides a new way to conduct PCA for sparse functional data. As shown in Section 4.1, this algorithm is more computationally stable than Yao, Müller and Wang (2005) by avoiding inverting the high-dimensional variance–covariance matrix. In terms of estimating component functions, the proposed method is comparable with the MLE method [Peng and Paul (2009)]. The difference between the proposed method and MLE methods is essentially the difference between least square regression and MLE estimator. However, the regression framework has its own advantages over the likelihood approaches. For example, one can replace the mean regressions with robust regressions when the data are contaminated with outliers. In addition, with minor modifications, the proposed regression based algorithm can also be used to conduct other types of functional decomposition such as singular value decomposition for functional data. By supporting various regression models and various decompositions, the proposed method can be extended to a rich family of lower dimension approximations for sparse functional data. Incorporating covariates and conducting variable selections are also straightforward under the regression framework. Our PCA algorithm estimates the mean and component functions nonparametrically. If there are additional recourses indicating certain parametric forms are more suitable, the efficiency of our method can be further improved.

SUPPLEMENTARY MATERIAL

Supplement to "Regression based principal component analysis for sparse functional data with applications to screening growth paths" (DOI: 10.1214/15-AOAS811SUPP; .zip). R programs for the proposed algorithm and an example of constructing the proposed growth chart.

REFERENCES

ABDOUS, B. and THEODORESCU, R. (1992). Note on the spatial quantile of a random vector. *Statist. Probab. Lett.* **13** 333–336. MR1160756

CHAKRABORTY, B. (2003). On multivariate quantile regression. *J. Statist. Plann. Inference* **110** 109–132. MR1944636

CHAUDHURI, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc.* **91** 862–872. MR1395753

CHEN, C., HE, X. and WEI, Y. (2008). Lower rank approximation of matrices based on fast and robust alternating regression. *J. Comput. Graph. Statist.* **17** 186–200. MR2424801

CHEN, K. and MÜLLER, H.-G. (2012). Conditional quantile analysis when covariates are functions, with application to growth data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 67–89. MR2885840

COLE, T. J. (1988). Fitting smoothed centile curves to reference data. *J. Roy. Statist. Soc. Ser. A* **151** 385–418.

COLE, T. J. and GREEN, P. J. (1992). Smoothing reference centile curves: The LMS method and penalized likelihood. *Stat. Med.* **11** 1305–1319.

CROUX, C., FILZMOSER, P., PISON, G. and ROUSSEEUW, P. J. (2003). Fitting multiplicative models by robust alternating regressions. *Stat. Comput.* **13** 23–36. MR1973864

DE BOOR, C. (1978). *A Practical Guide to Splines. Applied Mathematical Sciences* **27**. Springer, New York. MR0507062

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications. Monographs on Statistics and Applied Probability* **66**. Chapman & Hall, London. MR1383587

GRAVES, S., HOOKER, G. and RAMSAY, J. (2009). Functional data analysis with R and MATLAB.

HAN, B. and LIM, N. (2010). Estimating conditional proportion curves by regression residuals. *Stat. Med.* **29** 1443–1454. MR2758127

HANSEN, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* **24** 726–748. MR2409261

HETTMANSPERGER, T. P., NYBLOM, J. and OJA, H. (1992). On multivariate notions of sign and rank. In L_1*-Statistical Analysis and Related Methods* (*Neuchâtel*, 1992) 267–278. North-Holland, Amsterdam. MR1214838

JAMES, G. M., HASTIE, T. J. and SUGAR, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87** 587–602. MR1789811

KOLTCHINSKII, V. I. (1997). M-estimation, convexity and quantiles. *Ann. Statist.* **25** 435–477. MR1439309

LEGLER, J. D. and ROSE, L. C. (1998). Assessment of abnormal growth curves. *Am. Fam. Phys.* **58** 153–158.

LIU, R. Y., PARELIUS, J. M. and SINGH, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Ann. Statist.* **27** 783–858. MR1724033

LOEVE, M. (1978). Probability theory, Vol. II. *Grad. Texts in Math.* **46** 0–387.

MCDERMOTT, J. P. and LIN, D. K. (2007). Quantile contours and multivariate density estimation for massive datasets via sequential convex hull peeling. *IIE Trans.* **39** 581–591.

PARZEN, E. (1979). Nonparametric statistical data modeling. *J. Amer. Statist. Assoc.* **74** 105–131. MR0529528

PENG, J. and PAUL, D. (2009). A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *J. Comput. Graph. Statist.* **18** 995–1015. MR2598035

PENG, J. and PAUL, D. (2011). fpca: Restricted MLE for functional principal components analysis. R package version 0.2-1.

PERE, A. (2000). Comparison of two methods for transforming height and weight to normality. *Annals of Human Biology* **27** 35–45.

SCHEIKE, T. H., ZHANG, M.-J. and JUUL, A. (1999). Comparing reference charts. *Biom. J.* **41** 679–687.

SERFLING, R. (2002). Quantile functions for multivariate analysis: Approaches and applications. *Stat. Neerl.* **56** 214–232. MR1916321

THOMAS, G. B., FINNEY, R. L. and WEIR, M. D. (1988). *Calculus and Analytic Geometry* **7**. Addison-Wesley, Reading, MA.

THOMPSON, M. L. and FATTI, L. (1997). Construction of multivariate centile charts for longitudinal measurements. *Stat. Med.* **16** 333–345.

TREFETHEN, L. N. and BAU, D. III (1997). *Numerical Linear Algebra*. SIAM, Philadelphia, PA. MR1444820

WEI, Y. (2008). An approach to multivariate covariate-dependent quantile contours with application to bivariate conditional growth charts. *J. Amer. Statist. Assoc.* **103** 397–409. MR2420242

WEI, Y., PERE, A., KOENKER, R. and HE, X. (2006). Quantile regression methods for reference growth charts. *Stat. Med.* **25** 1369–1382. MR2226792

WOLD, H. (1966). Nonlinear estimation by iterative least square procedures. In *Research Papers in Statistics* (*Festschrift J. Neyman*) 411–444. Wiley, New York. MR0210250

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. MR2160561

ZHANG, W. (2012). Regression based principal component analysis for sparse functional data with applications to screening pubertal growth paths. Ph.D. thesis, Columbia Univ., New York.

ZHANG, W. and WEI, Y. (2015). Supplement to "Regression based principal component analysis for sparse functional data with applications to screening growth paths." DOI:10.1214/15-AOAS811SUPP.

ZUO, Y. and SERFLING, R. (2000). General notions of statistical depth function. *Ann. Statist.* **28** 461–482. MR1790005

DEPARTMENT OF BIOSTATISTICS
COLUMBIA UNIVERSITY
722 WEST 168TH ST. RM 644
NEW YORK, NEW YORK 10032
USA
E-MAIL: ying.wei@columbia.edu