# Semiparametric Modeling and Analysis of Longitudinal Method Comparison Data

**Lasitha N. Rathnayake and Pankaj K. Choudhary**[1]

Department of Mathematical Sciences, FO 35

University of Texas at Dallas

Richardson, TX 75083-0688, USA

**Abstract**

Studies comparing two or more methods of measuring a continuous variable are routinely conducted in biomedical disciplines with the primary goal of measuring agreement between the methods. Often, the data are collected by following a cohort of subjects over a period of time. This gives rise to longitudinal method comparison data where there is one observation trajectory for each method on every subject. It is not required that observations from all methods be available at each observation time. The multiple trajectories on the same subjects are dependent. We propose modeling the trajectories nonparametrically through penalized regression splines within the framework of mixed-effects models. The model also uses random effects of subjects and their interactions to capture dependence in observations from the same subjects. It additionally allows the within-subject errors of each method to be correlated. It is fit using the method of maximum likelihood. Agreement between the methods is evaluated by performing inference on measures of agreement, such as concordance correlation coefficient and total deviation index, which are functions of parameters of the assumed model. Simulations indicate that the proposed methodology performs reasonably well for thirty or more subjects. Its application is illustrated by analyzing a dataset of percentage body fat measurements.

[1]Corresponding author. Email: pankaj@utdallas.edu, Tel: (972) 883-4436, Fax: (972) 883-6622.

**Keywords**: Agreement, concordance correlation, mixed-effects model, longitudinal data, multiple outcome data, penalized splines, total deviation index.

# 1 Introduction

The usual longitudinal data arise when a cohort of subjects is followed over a period of time and observations of a response variable are recorded for each subject at various points in time. These data consist of one observation trajectory per subject. A key feature of the data is dependence in the observations from the same subjects. These data are commonly analyzed by modeling them in the framework of mixed-effects models which are fit by likelihood-based methods. See, e.g., [1, Chapter 9] and [2, Chapter 8] for introductions to this methodology. Multiple outcomes longitudinal data arise when observations of two or more response variables are recorded for each subject in the cohort. Thus, these data consist of multiple trajectories per subject. The observations from the same subjects are dependent but there are now two types of dependence — one for observations on the same trajectory and the other for observations on different trajectories. The analysis of these data calls for joint modeling of the multiple trajectories. One such model is developed in [3] for modeling of trajectories of diastolic and systolic blood pressure measurements of children as functions of their heights and weights.

Multiple outcomes longitudinal data are also of interest in this article. However, our interest is in the specific context of method comparison studies where the outcomes represent measurements from two or more methods of measuring a continuous variable that often has some clinical importance. The methods may be medical devices, clinical observers or assays. Such studies are common in medicine and other health-related disciplines [4, 5]. The main goal of a method comparison study is to evaluate whether the methods agree with each

other well enough to be used interchangeably. This would permit use of the cheapest or the most convenient method in place of the others. Agreement is quantified using *measures of agreement* such as concordance correlation coefficient (CCC, [6]) and total deviation index (TDI, [7–9]), which are functions of parameters of the joint distribution of the methods' measurements. The books [11, Chapters 2-4], [12, Chapters 2-6], and [13, Chapters 1-2], and the review article [14] can be consulted for an introduction to the analysis of method comparison data.

Statistical approaches for analysis of longitudinal method comparison data from two methods go back to [15]. This article assumes a random-coefficient growth curve model for the data and allows the within-subject variation to vary across subjects. It develops a CCC for each subject, and combines the subject-specific CCCs into a single index by taking their weighted average. The weights are based on the within-subject variation of the subjects. The articles of [16–18] consider the time of observation as a discrete quantity, measured as 'visit number' with, say $m$ possible values, and assume that each subject contributes two $m \times 1$ vectors of observations which are treated as draws from two multivariate populations. In [16], a 'repeated measures' CCC is developed that quantifies agreement in the two $m \times 1$ population vectors in a single index. An $m \times m$ weight matrix is also incorporated into the index to allow differential emphasis on within-visit and between-visit agreement. In [17], first an $m \times m$ matrix analog of CCC is developed by directly extending the CCC's definition to the multivariate case, and then its Frobenius norm is taken to get an overall measure of agreement. In [18], the data are modeled as a random-effects analysis of variance (ANOVA) model. Agreement is measured by an intraclass correlation coefficient defined under the model which is a special case of the repeated measures CCC developed in [16].

In case of longitudinal method comparison data with continuous observation times, the extent of agreement between the methods changes over time. However, none of the ap-

proaches of [15–18] allows this possibility as they all use a single overall index to measure agreement. This limitation is overcome in [19]. It takes differences in measurements of the methods, models their mean function nonparametrically using penalized regression splines in time, and measures agreement using TDI, which, under the model, is a function of time as well as the model parameters. This approach, however, has the drawback that it focuses on differences. This has two consequences. First, it does not use all the data, and in particular, throws away all those observations whose counterparts from the other methods are missing. Modeling of all data allows borrowing of information across methods and subjects which in turn helps in comparison of methods even if observations of some methods are missing at certain times or on certain subjects. Second, it precludes evaluating agreement using measures such as CCC, which require modeling of all data, not just the differences. The approaches of [16, 17] also discard observations that do not have counterparts from all the methods. Furthermore, all the existing approaches [15–19] share the drawbacks that they focus on specific agreement measures — [15–18] on CCC and [19] on TDI.

The aim of this article is to present a methodology that overcomes the aforementioned drawbacks of the existing approaches. We consider data from multiple methods, but it is not necessary that an observation from each method be available from every visit of a subject. We model the mean functions of the methods nonparametrically using separate penalized regression splines via their representation as a mixed-effects model. A lucid account of this kind of semiparametric modeling where the mean functions are specified nonparametrically but the model is actually fit as a parametric model is provided by [20, Chapters 3-5]. Our data model also incorporates random subject effects and their interactions to capture dependence in observations from the same subject. The within-subject errors of each method are allowed to be correlated. The proposed model generalizes the model of [18] by treating time as a continuous covariate. Once we have a model for the data, any measure of agreement can be

written in terms of the model parameters. The parameters are estimated by the method of maximum likelihood (ML), and bootstrap is employed to get bias and variance estimates that are used for computing relevant confidence intervals for the parameter functions of interest. Simultaneous confidence bands are constructed for the functions that depend on time.

The rest of this article is organized as follows. In Section 2, we describe the proposed methodology for modeling and analysis of longitudinal method comparison data. Its properties are evaluated in Section 3 through a simulation study. Its application is illustrated in Section 4 by analyzing the body fat data. Section 5 concludes with a discussion. Appendix A contains some technical details. All the computations here have been performed using the statistical software `R` [21].

# 2 Modeling and analysis of data

Suppose there are $J\,(\geq 2)$ methods under comparison, indexed as $j = 1, \ldots, J$. Suppose also that there are $n$ subjects in the study, indexed as $i = 1, \ldots, n$. We assume that the study plans for $m$ visits to collect data. Let $Y_{ij}(t_{ijk})$ denote the observation by the $j$th method on the $i$th subject at time $t_{ijk}$, $k = 1, \ldots, m_{ij}$, $j = 1, 2$, $i = 1, \ldots, n$. Here $t_{ijk}$ are values of the time covariate $t \in \mathcal{T}$, which is treated as a continuous variable. By definition, $m_{ij} \leq m$. The observations $Y_{ij}(t_{ijk})$, $k = 1, \ldots, m_{ij}$ form the trajectory of the $i$th subject under the $j$th method. Let $v_{ijk}$ denote the visit number during which the measurement $Y_{ij}(t_{ijk})$ taken. One can think of $v_{ijk}$ as the value of a discrete variable $v \in \{1, \ldots, m\}$. It is primarily used to link the observations taken during the same visit. Having observations from all methods during each visit is not necessary. The available observations, however, are linked by the common visit number. The index $v_{ijk}$ for it is generally different from the index $k$, representing the repeated measurement number. For example, the second repeated measurement may be

taken during the third visit, implying $k = 2$ and $v_{ij2} = 3$. One exception is when observations from all methods are available during each visit. In this case, $k$ and $v_{ijk}$ are identical.

Let $N_i = \sum_{j=1}^{J} m_{ij}$ be the total number of observations on the $i$th subject, and $N = \sum_{i=1}^{n} N_i$ be the total number of observations in the data. We use boldface letters to denote vectors and matrices. The transpose of $\mathbf{Y}$ is denoted as $\mathbf{Y}^T$. The vectors are column vectors unless specified otherwise. A $p$-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is denoted as $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

## 2.1 The data model

We propose modeling the observed data as a mixed-effects model,

$$Y_{ij}(t_{ijk}) = \mu_j(t_{ijk}) + b_i + b_{ij} + b^*_{iv_{ijk}} + e_{ijk},\ k = 1, \ldots, m_{ij},\ j = 1, \ldots, J,\ i = 1, \ldots, n, \quad (1)$$

where $\mu_j(t)$ is the mean function of the $j$th method, $b_i$ is the random effect of the $i$th subject, $b_{ij}$ is the random subject $\times$ method interaction, $b^*_{iv_{ijk}}$ is the random subject $\times$ visit number interaction, and $e_{ijk}$ is the within-subject random error. It is assumed that $b_i \sim$ independent $\mathcal{N}_1(0, \sigma_b^2)$, $b_{ij} \sim$ independent $\mathcal{N}_1(0, \psi^2)$ and $b^*_{iv_{ijk}} \sim$ independent $\mathcal{N}_1(0, \eta^2)$. For a given $(i, j)$ combination, the $e_{ijk}$ follow $\mathcal{N}_1(0, \sigma_{ej}^2)$ distributions with

$$\text{corr}(e_{ijk}, e_{ijl}) = \rho^{|t_{ijk} - t_{ijl}|}, \quad 0 \le \rho < 1.$$

This autocorrelation structure is a continuous time analog of AR(1), the autoregressive structure of order one. It is a simple model that allows unequally spaced times, and tends to work well in practice; see, e.g., [3] and the simulation results in Section 3. If needed, it can be replaced by other correlation models, see, e.g., spatial correlation models in [22, Chapter 5]. The errors for different $i$ or $j$ are assumed to be independent. The random effects and the errors in (1) are also assumed to be mutually independent. This model generalizes

the random-effects ANOVA model of [18] by treating time as a continuous covariate rather than as visit number, a discrete covariate.

The model assumptions for (1) imply that the observations of subject $i$ are jointly normally distributed with

$$E\{Y_{ij}(t)\} = \mu_j(t), \quad \text{var}\{Y_{ij}(t)\} = \sigma_b^2 + \psi^2 + \eta^2 + \sigma_{ej}^2. \tag{2}$$

For two observations on the same trajectory, we have

$$\text{cov}\{Y_{ij}(t_{ijk}), Y_{ij}(t_{ijl})\} = \sigma_b^2 + \psi^2 + \sigma_{ej}^2 \rho^{|t_{ijk} - t_{ijl}|}, \quad k \neq l. \tag{3}$$

This covariance structure depends on continuous time $t$ through the correlation in within-subject errors. For observations on different trajectories, depending upon whether or not they are collected in the same visit, we have

$$\text{cov}\{Y_{i1}(t_{i1k}), Y_{i2}(t_{i2l})\} = \begin{cases} \sigma_b^2 + \eta^2, & k = l, \\ \sigma_b^2, & k \neq l. \end{cases} \tag{4}$$

The observations from different subjects are mutually independent.

The mean functions in (1) are modeled nonparametrically via penalized splines regression. For this, we write $\mu_j(t)$ as a $p$th degree spline model,

$$\mu_j(t) = \beta_{0j} + \beta_{1j} t + \ldots + \beta_{pj} t^p + \sum_{q=1}^{Q} u_{qj} (t - c_q)_+^p, \quad j = 1, \ldots, J, \tag{5}$$

where $\beta_{0j}, \ldots, \beta_{pj}$ are regression coefficients, $Q$ is the number of knots, $c_1 < \ldots < c_Q$ are the locations of the knots, and $u_{1j}, \ldots, u_{Qj}$ are respective coefficients of the truncated polynomial basis functions $(t - c_1)_+^p, \ldots, (t - c_Q)_+^p$, with $(t - c_q)_+^p = (\max\{0, t - c_q\})^p$. The same set of knots are used for both mean functions. To estimate the coefficients, we adopt the mixed-effects model representation of the spline model wherein the spline coefficients $u_{qj}$ are treated as independent $\mathcal{N}_1(0, \sigma_{uj}^2)$ random variables [20, Chapters 4 and 6]. This makes the mean functions and any other parameter functions that depend on them random quantities.

Thus, the proposed model for the data is (1) with mean functions given by (5). The degree $p$, the number of knots $Q$, and the locations of the knots can be chosen according to the recommendations of [20, Chapter 5]. A matrix formulation of the model is given in (A.3) in Appendix A.

## 2.2 Evaluation of agreement

To evaluate agreement, we need longitudinal data analogs of measures of agreement under the assumed data model. For this, we first derive the joint distribution of $(\tilde{Y}_1(t), \ldots, \tilde{Y}_J(t))$, the observations taken by the $J$ methods on a randomly chosen subject from the underlying population at time $t \in \mathcal{T}$. The companion model for $\tilde{Y}_j(t)$ induced by the data model (1) is

$$\tilde{Y}_j(t) = \mu_j(t) + \tilde{b} + \tilde{b}_j + \tilde{b}^* + \tilde{e}_j, \quad j = 1, \ldots, J,$$

where $\tilde{b} \sim \mathcal{N}_1(0, \sigma_b^2)$, $\tilde{b}_j \sim$ independent $\mathcal{N}_1(0, \psi^2)$, $\tilde{b}^* \sim \mathcal{N}_1(0, \eta^2)$, $e_j \sim$ independent $\mathcal{N}_1(0, \sigma_{ej}^2)$, and the random variables are mutually independent. It follows that

$$\begin{bmatrix} \tilde{Y}_1(t) \\ \tilde{Y}_2(t) \\ \vdots \\ \tilde{Y}_J(t) \end{bmatrix} \sim \mathcal{N}_J\left( \begin{bmatrix} \mu_1(t) \\ \mu_2(t) \\ \vdots \\ \mu_J(t) \end{bmatrix}, \begin{bmatrix} \sigma_b^2 + \psi^2 + \eta^2 + \sigma_{e1}^2 & \sigma_b^2 + \eta^2 & \ldots & \sigma_b^2 + \eta^2 \\ \sigma_b^2 + \eta^2 & \sigma_b^2 + \psi^2 + \eta^2 + \sigma_{e2}^2 & \ldots & \sigma_b^2 + \eta^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 + \eta^2 & \sigma_b^2 + \eta^2 & \ldots & \sigma_b^2 + \psi^2 + \eta^2 + \sigma_{eJ}^2 \end{bmatrix} \right). \tag{6}$$

Two common measures of agreement between a pair of methods are CCC [6] and TDI [7]. For measuring agreement between methods $j$ and $r$, they can be expressed using (6) as

$$\begin{aligned} \mathrm{CCC}_{jr}(t) &= \frac{2\mathrm{cov}\{\tilde{Y}_j(t), \tilde{Y}_r(t)\}}{[E\{\tilde{Y}_j(t)\} - E\{\tilde{Y}_r(t)\}]^2 + \mathrm{var}\{\tilde{Y}_j(t)\} + \mathrm{var}\{\tilde{Y}_r(t)\}} \\ &= \frac{2(\sigma_b^2 + \eta^2)}{\{\mu_j(t) - \mu_r(t)\}^2 + 2(\sigma_b^2 + \psi^2 + \eta^2) + \sigma_{ej}^2 + \sigma_{er}^2} \end{aligned} \tag{7}$$

and

$$\begin{aligned}\mathrm{TDI}_{jr}(p_0,t) &= 100p_0\text{th percentile of } |\tilde{Y}_j(t)-\tilde{Y}_r(t)| \\ &= \left\{(2\psi^2+\sigma_{ej}^2+\sigma_{er}^2)\,\chi^2_{1,p_0}\left(\frac{\{\mu_j(t)-\mu_r(t)\}^2}{2\psi^2+\sigma_{ej}^2+\sigma_{er}^2}\right)\right\}^{1/2}, \end{aligned} \tag{8}$$

where $p_0$ is a specified large probability and $\chi^2_{1,p_0}(\Delta)$ represents the $100p_0$th percentile of a noncentral $\chi^2$ distribution with one degree of freedom and noncentrality parameter $\Delta$.

The CCC lies between zero and one, and a large value for it implies good agreement. It equals $\mathrm{corr}\{\tilde{Y}_j(t),\tilde{Y}_r(t)\}$ if $\mu_j(t)=\mu_r(t)$ and $\sigma_{ej}^2=\sigma_{er}^2$, and is zero if the methods are uncorrelated. A CCC is related to an intraclass correlation, see, e.g., [14]. The TDI is bounded below by zero, and a small value for it implies good agreement. Perfect agreement at $t$, the case when the bivariate distribution of $(\tilde{Y}_j(t),\tilde{Y}_r(t))$ is concentrated on the 45° line, is implied when the CCC equals one or equivalently the TDI equals zero. We evaluate agreement by examining one-sided simultaneous confidence bands for agreement measures over $\mathcal{T}$, in particular, a lower band for CCC and an upper band for TDI, for each pair of methods of interest. Their construction is described in the following subsection.

## 2.3 Inference under the model

Let $\boldsymbol{\theta}$ be the vector of unknown parameters in the assumed data model. We obtain its ML estimator $\hat{\boldsymbol{\theta}}$ by numerically maximizing the likelihood under (A.3). Any mainstream software package for fitting mixed-effects models, e.g., the `nlme` package [23] in `R`, can be used for this purpose. The functions of $\boldsymbol{\theta}$ are estimated by plugging in $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$. From (5), the fitted mean functions are

$$\hat{\mu}_j(t)=\hat{\beta}_{0j}+\hat{\beta}_{1j}t+\ldots+\hat{\beta}_{pj}t^p+\sum_{q=1}^{Q}\hat{u}_{qj}(t-c_q)_+^p,\quad j=1,\ldots,J, \tag{9}$$

where the $\hat{u}_{qj}$ are the estimated best linear unbiased predictors (BLUPs; see [22, Chapter 2] and Appendix A).

We can get the standard errors of ML estimators of components of $\boldsymbol{\theta}$ and their confidence intervals in the usual manner via the large-sample theory of ML estimators [24]. However, this theory does not directly apply for inference on the mean functions $\mu_j(t)$ and the parameter functions that involve them because these are random quantities under the mixed-effects model representation of the spline functions, see (5) and also [20, Chapters 4 and 6]. We deal with this difficulty by using parametric bootstrap. The bootstrap readily yields estimates of biases and variances of the estimators, which are then combined with the assumption of approximate normality of the estimators to get simultaneous confidence bands for the parameter functions.

To fix ideas, suppose $\phi(t) \equiv \phi(t, \boldsymbol{\theta}, \mathbf{u})$, $t \in \mathcal{T}$ is a function of parameter vector $\boldsymbol{\theta}$ and spline coefficients vector $\mathbf{u}$ (see Appendix A). Specific examples of $\phi(t)$ include the mean function $\mu_j(t)$, the mean difference function $\mu_r(t) - \mu_j(t)$, and the agreement measures $\mathrm{CCC}_{jr}(t)$ and $\mathrm{TDI}_{jr}(p_0, t)$ given by (7) and (8), respectively. We would like to construct an appropriate one- or two-sided simultaneous confidence band for $\phi(t)$ over $\mathcal{T}$. The band is actually constructed over a relatively fine grid $\mathbf{t} = (t_1, \ldots, t_L)^T$ of $L$ points in $\mathcal{T}$. In practice, $L$ between 25 to 50 is generally adequate. The estimator $\hat{\phi}(t) \equiv \phi(t, \hat{\boldsymbol{\theta}}, \hat{\mathbf{u}})$ of $\phi(t)$ is obtained by replacing $\boldsymbol{\theta}$ by its estimator $\hat{\boldsymbol{\theta}}$ and $\mathbf{u}$ by its estimated BLUP $\hat{\mathbf{u}}$. Let $\hat{\phi}(\mathbf{t})$ and $\phi(\mathbf{t})$ respectively represent the $L \times 1$ vectors of values of $\hat{\phi}(t)$ and $\phi(t)$ on the grid. Owing to the randomness in $\phi(t)$, we need the estimated mean vector and covariance matrix of $\hat{\phi}(\mathbf{t}) - \phi(\mathbf{t})$ to construct the band. Let these be denoted by the $L \times 1$ vector $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, \ldots, \hat{\delta}_Q)^T$ and the $L \times L$ matrix $\hat{\mathbf{S}} = (\hat{s}_{lr})_{l,r=1,\ldots L}$, respectively. Upon approximating the distribution of $\hat{\phi}(\mathbf{t}) - \phi(\mathbf{t})$ for large $n$ by a $\mathcal{N}_L(\hat{\boldsymbol{\delta}}, \hat{\mathbf{S}})$ distribution, possibly after applying a normalizing transformation,

an approximate $100(1-\alpha)\%$ simultaneous confidence band for $\boldsymbol{\phi}(\mathbf{t})$ can be computed as:

$$\text{lower band: } \hat{\phi}(t_l) - \hat{\delta}_l - a_{1-\alpha,L}\sqrt{\hat{s}_{ll}}, \quad \text{upper band: } \hat{\phi}(t_l) - \hat{\delta}_l + a_{1-\alpha,L}\sqrt{\hat{s}_{ll}},$$
$$\text{two-sided band: } \hat{\phi}(t_l) - \hat{\delta}_l \pm b_{1-\alpha,L}\sqrt{\hat{s}_{ll}}, \quad l = 1, \ldots, L.$$

The critical point $a_{1-\alpha,L}$ is the $100(1-\alpha)$th percentile of the maximum of the components of a $L$-variate normal vector with mean zero and covariance matrix equal to the correlation matrix corresponding to $\hat{\mathbf{S}}$. The critical point $b_{1-\alpha,L}$ is defined similarly except that the maximum of the absolute values of the components is taken. The critical points ensure that the simultaneous coverage probability of the band is approximately $1-\alpha$. They can be computed using the `multcomp` package of [25] in `R`. When $L = 1$, $a_{1-\alpha,L} = z_{1-\alpha}$ and $b_{1-\alpha,L} = z_{1-\alpha/2}$, where $z_\alpha$ represents the $100\alpha$th percentile of a $\mathcal{N}_1(0,1)$ distribution.

Now all that remains is the estimation of the bias vector $\hat{\boldsymbol{\delta}}$ and the covariance matrix $\hat{\mathbf{S}}$. Below we present a parametric bootstrap methodology for this. It uses the matrix formulation of the model (A.3) presented in Appendix A and involves the following steps:

1. Simulate $\mathbf{u}^* \sim \mathcal{N}_{JQ}(\mathbf{0}, \hat{\mathbf{G}}_\mathbf{u})$, $\mathbf{b}^* \sim \mathcal{N}_{n(1+J+m)}(\mathbf{0}, \text{diag}\{\hat{\mathbf{G}}_{\mathbf{b}_1}, \ldots, \hat{\mathbf{G}}_{\mathbf{b}_n}\})$ and also $\mathbf{e}^* \sim \mathcal{N}_N(\mathbf{0}, \text{diag}\{\hat{\mathbf{R}}_1, \ldots, \hat{\mathbf{R}}_n\})$. Follow (A.3) to form

$$\mathbf{Y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{W}\mathbf{u}^* + \mathbf{Z}\mathbf{b}^* + \mathbf{e}^*$$

   as a bootstrap resample of the original sample $\mathbf{Y}$.

2. Compute $\phi^*(\mathbf{t}) \equiv \phi(\mathbf{t}, \hat{\boldsymbol{\theta}}, \mathbf{u}^*)$ as the bootstrap counterpart of $\phi(\mathbf{t})$.

3. Fit model (A.3) to the resampled data $\mathbf{Y}^*$ by ML and get $\hat{\boldsymbol{\theta}}^*$ as the resulting estimate of $\boldsymbol{\theta}$.

4. Compute $\hat{\mathbf{u}}^*$, the bootstrap counterpart of the estimated BLUP $\hat{\mathbf{u}}$, by replacing $(\mathbf{Y}, \hat{\boldsymbol{\theta}})$ in its expression with $(\mathbf{Y}^*, \hat{\boldsymbol{\theta}}^*)$.

5. Compute $\hat{\phi}^*(\mathbf{t}) \equiv \phi(\mathbf{t}, \hat{\boldsymbol{\theta}}^*, \hat{\mathbf{u}}^*)$ as the bootstrap counterpart of $\hat{\phi}(\mathbf{t})$, and get the difference $\hat{\phi}^*(\mathbf{t}) - \phi^*(\mathbf{t})$.

6. Repeat Steps 1-5 a large number of times, say $B$. Denote the difference $\hat{\phi}^*(\mathbf{t}) - \phi^*(\mathbf{t})$ computed in the $b$th repetition as $\hat{\phi}_l^*(\mathbf{t}) - \phi_l^*(\mathbf{t})$, $l = 1, \ldots, B$.

7. Take $\hat{\boldsymbol{\delta}}$ and $\hat{\mathbf{S}}$ as respectively the sample mean and the sample covariance matrix of the $B$ differences in Step 6.

Typically $B = 500$ is adequate for the task of estimating bias and covariance matrix. Although we have taken $\phi(t)$ to be a parameter function involving $(t, \boldsymbol{\theta}, \mathbf{u})$, the bootstrap method can also be used for inference on elements of $\boldsymbol{\theta}$ alone or functions thereof, e.g., the precision ratio $\sigma_{e1}^2/\sigma_{e2}^2$, by suitably modifying the foregoing algorithm. This is the approach we take here. Also, for increased accuracy of the normal approximation, the confidence intervals are first computed by applying a normalizing transformation to the parameter to make its range $(-\infty, \infty)$, and then transforming the results back to the original scale. In particular, a Fisher's $z$-transformation is applied to CCC. It is defined as

$$z(\text{CCC}) = \frac{1}{2} \log \left( \frac{1 + \text{CCC}}{1 - \text{CCC}} \right) = \tanh^{-1}(\text{CCC}).$$

Further, a log transformation is applied to standard deviations, precision ratio and TDI, and a logit transformation is applied to the autocorrelation parameter $\rho$.

So far in this section we have assumed the data model (1). In practice, sometimes it may happen that the data support only a special case of the model, e.g., the model without the subject $\times$ method interaction or the model with independent errors. A real example of this situation is provided by the body fat data in Section 4. The proposed methodology can be easily adapted to deal with such cases. This involves reformulating the new model in the matrix form (A.3) as described in Appendix A, redefining $\boldsymbol{\theta}$ appropriately, and if necessary, obtaining expressions for the agreement measures (7) and (8) under the new model.

# 3 A Simulation Study

This section is devoted to evaluation of the performance of the methodology proposed in the previous section for $J = 2$ methods via Monte Carlo simulation. We are specifically interested in examining the accuracy of the individual confidence interval for $\sigma_{e1}^2/\sigma_{e2}^2$, simultaneous two-sided confidence bands for $\mu_1(t)$, $\mu_2(t)$ and $\mu_2(t) - \mu_1(t)$, simultaneous lower confidence band for CCC$(t)$, and simultaneous upper confidence band for TDI$(p_0, t)$, for $t \in \mathcal{T}$. Also of interest is a study of how this accuracy is affected if the true autocorrelation structure of the within-subject errors in model (1) differs from the assumed continuous AR(1) structure.

The data are simulated from model (1) along the lines of the body fat data but to allow moderate agreement between the measurement methods — a common scenario in practice. Specifically, we take $\mathcal{T} = (10.5, 16)$, mean functions $\mu_1(t) = -325 + 70t - 2.5t^2$, $\mu_2(t) = 575 - 70t + 2.5t^2$, variance components $(\sigma_b^2, \psi^2, \eta^2, \sigma_{e1}^2, \sigma_{e2}^2) = (230, 20, 15, 120, 90)$, and seven models for autocorrelation in within-subject errors. They are: independent errors; AR(1) with autocorrelation parameter equal to 0.2 and 0.5; MA(1), i.e., moving average model of order one, with autocorrelation parameter equal to 0.2 and 0.5; and compound symmetric model with common correlation equal to 0.2 and 0.5. These models as respectively denoted as IND, AR(0.2), AR(0.5), MA(0.2), MA(0.5), CS(0.2) and CS(0.5). We simulate two trajectories per subject for $n \in \{30, 60, 100\}$ subjects using both balanced and unbalanced designs. For the balanced design, each subject has nine visits, starting randomly around age 11 and being six months apart on average. Observations from both methods are available in each visit. Thus, in this case, each trajectory of a subject has nine observations. The data for the unbalanced design are initially simulated as in the balanced design but only the observations from a Poisson (6) number of visits (up to a maximum of nine visits), simulated independently for each method, are randomly selected to be kept in the data, and

the rest are set to be missing. Thus, in this case, the two trajectories of a subject may not have an equal numbers of observations, and it is not necessary that the observations from both methods are available in a particular visit. Altogether we have $7 * 3 * 2 = 42$ settings.

For each setting, we simulate the data as explained above, fit model (1) to the data assuming the continuous AR(1) structure regardless of the true autocorrelation model, and use the methodology of Section 2 with $B = 500$ bootstrap replications to compute the desired individual confidence intervals and simultaneous confidence bands with $1 - \alpha = 0.95$. The simultaneous bands employ a grid of $L = 30$ equally spaced points on $\mathcal{T}$. The probability $p_0$ for TDI is taken to be 0.9. The process of simulating data and constructing confidence intervals and bands is repeated 500 times, and an appropriate coverage probability — either individual or simultaneous — is computed for each interval and band.

Table 1 presents the estimated individual coverage probabilities for the confidence interval for $\sigma_{e1}^2/\sigma_{e2}^2$ and simultaneous coverage probabilities for the confidence bands for $\mu_1(t)$, $\mu_2(t)$, $\mu_2(t) - \mu_1(t)$, CCC$(t)$ and TDI$(0.90, t)$ for the balanced design. Similar results for the unbalanced design are presented in Table 2. In the balanced case, we see that with the exception of CCC, all coverage probabilities are close to the nominal level of 95%. The CCC bands appear a bit conservative as their coverage probabilities are 2-3% greater than the nominal level. Interestingly, there is no considerable difference in the probabilities across the various autocorrelation models. This indicates that the methodology is robust to the autocorrelation structure in the data. The results for $n = 30, 60$ and 100 also appear remarkably similar. The results in case of the unbalanced design follow the same pattern as the balanced design. In some instances, probabilities for $n = 60, 100$ are closer to the nominal level than for $n = 30$. This is especially true for the CCC band. These conclusions are representative and are not overly tied to the specific parameter values chosen. Overall, they suggest that the proposed methodology has acceptable accuracy with $n = 30$ for both balanced and un-

balanced designs. Moreover, the true autocorrelation structure of the within-subject errors does not seem to have much impact on the accuracy of the methodology.

# 4 Application: Body Fat Data

These data are from [15] and have also been analyzed in [16, 17, 19]. They consist of percentage body fat measurements estimated using two methods, namely, skinfold calipers and dual energy x-ray absorptiometry (DEXA), on a cohort of 112 girls over a period of about $4\frac{1}{2}$ years. The first visit occurred around age twelve and there were eight subsequent visits roughly six months apart. The actual ages of the girls are also recorded at each visit. We do not have observations from both methods on each visit. In particular, the DEXA observations are completely missing from the first visit, and there are 75 girls for whom observations from one or both methods are missing from at least one of the later visits. In total, there are 1,515 observations in the data, ranging from 12.7 to 37.4 with an average of 23.5. The ages range from 10.7 to 17.3 years with an average of 13.8 years. Some lack of agreement between the two methods is expected given that they estimate percentage body fat quite differently. For example, the measurements of skinfold calipers are based on measurements of thicknesses of subcutaneous fat layer at several standardized points in the body. On the other hand, DEXA measurements are obtained by scanning the body with two x-ray beams with different energy levels and examining the difference in their absorptions by soft tissues. The main goal here is to see if the methods agree sufficiently well to be used interchangeably.

Figure 1 displays the body fat trajectories separately for caliper and DEXA methods. Also superimposed on the plots are the mean functions estimated using the methodology proposed in Section 2 (see below for details). There is considerable variation in the tra-

jectories both within and between the methods. Further, the two mean functions change with age, and the same holds for their difference. This implies that the extent of agreement between the methods changes with age.

To analyze these data, the skinfold calipers and DEXA methods are taken to be methods 1 and 2, respectively. The time covariate $t$ is the age (in years) of the girl at the time of measurement. The observed age range of $(10.7, 17.3)$ years serves as the interval $\mathcal{T}$. Our first task is to model the data. Initially, we fit the model (1) with $J = 2$ and mean functions (5). The degree of the spline model is chosen to be $p = 2$ to impart a smooth appearance to the estimated mean functions. This choice, however, is not crucial for our purposes as similar results are obtained, e.g., with $p = 1$ or 3. Following [20, Chapter 5], we take $Q = 35$ knots, located at $(1/36), \ldots, (35/36)$th sample quantiles of the observed unique time values. The likelihood ratio test of zero autocorrelation in the within-subject errors has a $p$-value of less than $10^{-3}$, confirming the need for incorporating the autocorrelation. However, then the random subject $\times$ method interaction, which from (3) also contributes to correlation between observations on the same trajectory, becomes redundant. The $p$-value of the relevant likelihood ratio test is 0.18. Therefore, we drop this term from the model, and assume the model without the interaction for all the inferences henceforth. The key conclusions, nevertheless, remain unchanged if the interaction is kept in the model. For model evaluation, [22, Chapters 4 and 5] suggests examining residual plots for assessing overall model fit, normal quantile-quantile plots for assessing normality of within-subject errors and random subject effects, and plots of empirical autocorrelation and semivariogram functions for assessing the continuous AR(1) autocorrelation structure for the errors. These diagnostics (not presented here) show that the assumed model provides a reasonably good fit to the data.

Table 3 provides estimates of the variance-covariance related parameters in the assumed

model. Their 95% confidence intervals are also provided. None of the intervals appears unusually wide. Substituting the ML estimates in the counterpart of (6) for the assumed model gives the fitted distribution of $(\tilde{Y}_1(t), \tilde{Y}_2(t))$, $t \in \mathcal{T}$ as

$$\begin{bmatrix} \tilde{Y}_1(t) \\ \tilde{Y}_2(t) \end{bmatrix} \sim \mathcal{N}_2 \left( \begin{bmatrix} \hat{\mu}_1(t) \\ \hat{\mu}_2(t) \end{bmatrix}, \begin{bmatrix} 15.3 & 9.8 \\ 9.8 & 14.0 \end{bmatrix} \right). \tag{10}$$

The estimated mean functions $\hat{\mu}_1(t)$ and $\hat{\mu}_2(t)$ are given by (9). They are plotted in Figure 1. The two functions do not have the same shape. Caliper's mean ranges between 22 to 27, and it exceeds DEXA's mean, which ranges between 21 to 25, for most of $\mathcal{T}$. Caliper's variance also exceeds that of DEXA's. The correlation of 0.67 between the methods cannot be considered high compared to what we normally see in method comparison studies.

Our next task is to evaluate agreement between the methods. The simultaneous confidence bands for measures that depend on $t$ are computed over a grid of thirty equally-spaced points in $\mathcal{T}$. The critical points needed for the bands are computed using the `multcomp` package of [25] in `R`. Figure 2 presents estimates and 95% simultaneous one-sided confidence bands for CCC and TDI with $p_0 = 0.90$. A lower band is presented for CCC and an upper band is presented for TDI. The critical points used in the two bands are $-2.55$ and 2.68, respectively. It is clear from the figure that both measures suggest the same pattern of agreement between the methods. The CCC estimate ranges between 0.5 to 0.7 and its lower bound ranges between 0.2 to 0.6. Likewise, the TDI estimate ranges between 4.7 to 7.3 and its upper bound ranges between 5.5 to 11.2. Based on the bounds, the agreement appears best around age 11 and between ages of 14.5 to 16.5, and worst around age 13 and near the endpoints. Specifically, the extent of agreement increases till about age 11, decreases sharply thereafter till about age 13.5, increases again till about age 14.5 and reaches its previous highest value, remains constant there till about age 16.5, and finally decreases again. The uncertainty in the bands is highest near the endpoints. In the best case scenario, CCC lower bound is 0.6

and the TDI upper bound is 5.5. The latter implies that 90% of differences in observations from the two methods fall within $\pm 5.5$. Given that the body fat measurements themselves average around 23.5, the level of agreement implied by TDI even in the best case scenario is unlikely to be considered high enough for interchangeable use of the methods. The same conclusion is reached on the basis of CCC as well. Thus, on the whole, the caliper and the DEXA methods do not agree well enough to be used interchangeably. These conclusions are consistent with [15, 16], where the authors also provide a physiological explanation for the especially poor agreement around age 13 by noting that "skinfold estimation is only capable of detecting subcutaneous fat, whereas breast, lower body and visceral fat are increasing over this age range due to the onset of menarche."

To gain some insight into the lack of agreement, we examine the estimate of difference in means of DEXA and caliper methods and its 95% simultaneous two-sided confidence band. These are presented in Figure 3. The critical point used in the band is 2.83. We see that the estimate is positive for ages below 11.3 years and between 15 to 16 years, and it is negative elsewhere. The confidence band is widest near the endpoints. It contains zero except when the age is between 11.8 to 14.4 years in which case the band is below zero. This implies that the means of the two methods cannot be considered equal over the entire age interval $\mathcal{T}$. The pattern of increase and decrease in absolute value of the estimated mean difference is consistent with the pattern of decrease and increase in the extent of agreement. Specifically, the maximum absolute difference is about 3.4, which occurs around age 13 years — the region where the estimated agreement is weakest. Thus, the lack of agreement is partly explained by a difference in the means of the methods. In addition, the precisions of the methods are relatively low, leading to a somewhat low correlation between them. The estimate of precision ratio of DEXA and caliper methods is 1.28 and its 95% confidence interval is $(1.04, 1.58)$. This implies that the DEXA method is more precise of the two.

Based on this, we may prefer the DEXA method because its precision is higher than that of the caliper method.

# 5 Discussion

In this article, we presented a semiparametric procedure for modeling longitudinal method comparison data from two or more measurement methods using penalized regression splines within the framework of mixed-effects models. The assumed model is then used to develop a methodology for evaluation of agreement between the methods. Our approach is flexible in that it can accommodate balanced or unbalanced data designs and multiple methods. Moreover, it does not require one to specify in advance how the mean functions may be related to time. Instead, these relationships are completely determined from the data. The approach also allows the agreement measures to depend on time to capture changes in the extent of agreement over time. The simulations indicate that the methodology has acceptable performance for 30 or more subjects, and is robust to autocorrelation structure of the within-subject errors. Although the proposed model assumed homoscedastic errors, it can be easily extended to deal with heteroscedastic errors along the lines of [26].

One limitation of the methodology is that it is computationally intensive. However, the computations can be easily programmed. The `R` program used to do the computations for the body fat data is available at the second author's professional website, `http://www.utdallas.edu/~pankaj`. The computations took about 200 minutes on a Windows computer with a 2.2 GHz processor and 4 GB RAM.

Here we have used the ML method for parameter estimation. The method of restricted maximum likelihood (REML) is another common method for parameter estimation in mixed-effects models. Ordinarily, the REML method has limited usefulness for inference on agree-

ment measures, which are functions of both mean and variance-covariance related parameters, because it does not provide a joint covariance matrix for estimators of all such parameters. However, this covariance matrix is not needed in the proposed methodology because bootstrap is used to estimate the uncertainty in the estimators. This implies that the REML method can also be used in place of the ML method.

## Acknowledgements

The authors thank Professor Vernon Chinchilli for providing the body fat data, and the Texas Advanced Computing Center at The University of Texas at Austin for providing HPC resources for conducting the simulation studies. We also thank two anonymous reviewers and the associate editor for providing constructive comments that have greatly improved this article.

## References

1. Diggle, P. J., Heagertry, P., Liang, K.-L. and Zeger, S. L. *Analysis of Longitudinal Data*. 2nd edn. Oxford University Press: New York, 2002.

2. Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis*. 2nd edn. John Wiley: Hoboken, New Jersey, 2011.

3. Liu H, Tu W. A semiparametric regression model for paired longitudinal outcomes with application in childhood blood pressure development. *The Annals of Applied Statistics* 2012; **6**:1861–1882.

4. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **i**:307–310.

5. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 1999; **8**:135–160.

6. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; **45**:255–268. Corrections: 2000, **56**:324-325.

7. Lin LI. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine* 2000; **19**:255–270.

8. Lin LI, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: Models, issues, and tools. *Journal of the American Statistical Association* 2002; **97**:257–270.

9. Choudhary PK, Nagaraja HN. Tests for assessment of agreement using probability criteria. *Journal of Statistical Planning and Inference* 2007; **137**:279–290.

10. Dunn G. Regression models for method comparison data. *Journal of Biopharmaceutical Statistics* 2007; **17**:739–756.

11. Dunn G. *Statistical Evaluation of Measurement Errors*. 2nd edn. John Wiley, Chichester, UK, 2004.

12. Carstensen B. *Comparing Clinical Measurement Methods: A Practical Guide*. John Wiley: New York, 2010.

13. Lin LI, Hedayat AS, Wu W. *Statistical Tools for Measuring Agreement*. Springer: New York, 2011.

14. Barnhart HX, Haber MJ, Lin LI. An overview on assessing agreement with continuous measurement. *Journal of Biopharmaceutical Statistics* 2007; **17**:529–569.

15. Chinchilli VM, Martel JK, Kumanyika S, Lloyd T. A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics* 1996; **52**:341–353.

16. King TS, Chinchilli VM, Carrasco JL. A repeated measures concordance correlation coefficient. *Statistics in Medicine* 2007; **26**:3095–3113.

17. Hiriote S, Chinchilli VM. Matrix-based concordance correlation coefficient for repeated measures. *Biometrics* 2011; **67**:1007–1016.

18. Carrasco JL, King TS, Chinchilli VM. The concordance correlation coefficient for repeated measures estimated by variance components. *Journal of Biopharmaceutical Statistics* 2009; **19**:90–105.

19. Choudhary PK. Semiparametric regression for assessing agreement using tolerance bands. *Computational Statistics and Data Analysis* 2007; **51**:6229–6241.

20. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Cambridge University Press: New York, 2003.

21. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org`.

22. Pinheiro JC, Bates DM. *Mixed-Effects Models in S and S-PLUS*. Springer: New York, 2000.

23. Pinheiro JC, Bates D, DebRoy S, Sarkar D, R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2015. URL `http://CRAN.R-project.org/package=nlme`.

24. Lehmann EL. *Elements of Large-Sample Theory*. Springer: New York, 1998.

25. Hothorn T, Bretz F, Westfall P. Simultaneous inference in general parametric models. *Biometrical Journal* 2008; **50**:346–363.

26. Nawarathna LS, Choudhary PK. Measuring agreement in method comparison studies with heteroscedastic measurements. *Statistics in Medicine* 2013; **32**:5156–5171.

27. Robinson GK. That BLUP is a good thing: The estimation of random effects. *Statistical Science* 1991; **6**:15–32.

## Appendix A. Model in matrix notation

In this section, we formulate the model (1) with mean functions (5) using matrix notation. For $i = 1, \ldots, n,\ j = 1, \ldots, J$, define $\mathbf{Y}_{ij} = (Y_{ij}(t_{ij1}), \ldots, Y_{ij}(t_{ijm_{ij}}))^T$, $\mathbf{e}_{ij} = (e_{ij1}, \ldots, e_{ijm_{ij}})^T$, $\boldsymbol{\beta}_j = (\beta_{0j}, \ldots, \beta_{pj})^T$,

$$\mathbf{X}_{ij} = \begin{bmatrix} 1 & t_{ij1} & \ldots & t_{ij1}^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{ijm_{ij}} & \ldots & t_{ijm_{ij}}^p \end{bmatrix},$$

and

$$\mathbf{Y}_i = \begin{bmatrix} \mathbf{Y}_{i1} \\ \vdots \\ \mathbf{Y}_{iJ} \end{bmatrix},\ \mathbf{e}_i = \begin{bmatrix} \mathbf{e}_{i1} \\ \vdots \\ \mathbf{e}_{iJ} \end{bmatrix},\ \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_J \end{bmatrix},\ \mathbf{X}_i = \begin{bmatrix} \mathbf{X}_{i1} & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \mathbf{X}_{iJ} \end{bmatrix}.$$

Next, define $\mathbf{u}_j = (u_{1j}, \ldots, u_{Qj})^T$,

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_J \end{bmatrix},\ \mathbf{W}_{ij} = \begin{bmatrix} (t_{ij1} - c_1)_+^p & \ldots & (t_{ij1} - c_Q)_+^p \\ \vdots & \vdots & \vdots \\ (t_{ijm_{ij}} - c_1)_+^p & \ldots & (t_{ijm_{ij}} - c_Q)_+^p \end{bmatrix},\ \mathbf{W}_i = \begin{bmatrix} \mathbf{W}_{i1} & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \mathbf{W}_{iJ} \end{bmatrix},$$

and also

$$\mathbf{b}_i = \begin{bmatrix} b_i \\ b_{i1} \\ \vdots \\ b_{iJ} \\ b^*_{i1} \\ \vdots \\ b^*_{im} \end{bmatrix}, \ \mathbf{Z}_i = \begin{bmatrix} 1 & 1 & \dots & 0 & I(v_{i11}=1) & \dots & I(v_{i11}=m) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \dots & 0 & I(v_{i1m_{i1}}=1) & \dots & I(v_{i1m_{i1}}=m) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 1 & I(v_{iJ1}=1) & \dots & I(v_{iJ1}=m) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 1 & I(v_{iJm_{iJ}}=1) & \dots & I(v_{iJm_{iJ}}=m) \end{bmatrix}.$$

Now the model (1) with mean functions (5) can be written as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{u} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \ i = 1, \dots, n, \tag{A.1}$$

where

$\mathbf{u} \sim \mathcal{N}_{JQ}(\mathbf{0}, \mathbf{G}_\mathbf{u}), \ \mathbf{G}_\mathbf{u} = \text{diag}\{\mathbf{G}_{\mathbf{u}_1}, \dots, \mathbf{G}_{\mathbf{u}_J}\}, \ \mathbf{G}_{\mathbf{u}_j} = \text{diag}\{\sigma^2_{uj}, \dots, \sigma^2_{uj}\}, \ j = 1, \dots, J,$

$\mathbf{b}_i \sim \text{independent } \mathcal{N}_{(1+J+m)}(\mathbf{0}, \mathbf{G}_{\mathbf{b}_i}), \ \mathbf{G}_{\mathbf{b}_i} = \text{diag}\{\sigma^2_b, \psi^2, \dots, \psi^2, \eta^2, \dots, \eta^2\},$

$\mathbf{e}_i \sim \text{independent } \mathcal{N}_{N_i}(\mathbf{0}, \mathbf{R}_i), \ \mathbf{R}_i = \text{diag}\{\mathbf{R}_{i1}, \dots, \mathbf{R}_{iJ}\},$

with $\mathbf{R}_{ij}$ as an $m_{ij} \times m_{ij}$ covariance matrix whose $(k,l)$th element is $\sigma^2_{ej}\rho^{|t_{ijk} - t_{ijl}|}$. The vectors $\mathbf{u}$, $\mathbf{b}_i$ and $\mathbf{e}_i$ are mutually independent. It follows that $\mathbf{Y}_i \sim \mathcal{N}_{N_i}(\mathbf{X}_i\boldsymbol{\beta}, \text{var}(\mathbf{Y}_i))$ with

$$\text{var}(\mathbf{Y}_i) = \mathbf{W}_i\mathbf{G}_\mathbf{u}\mathbf{W}_i^T + \mathbf{Z}_i\mathbf{G}_{\mathbf{b}_i}\mathbf{Z}_i^T + \mathbf{R}_i. \tag{A.2}$$

The $\mathbf{Y}_i$ are not independent because they share the common spline coefficient vectors $\mathbf{u}$. We can see that

$$\text{cov}(\mathbf{Y}_i, \mathbf{Y}_l) = \mathbf{W}_i\mathbf{G}_\mathbf{u}\mathbf{W}_l^T, \ i \neq l.$$

To obtain a joint model for all the $N$ observations in the data, define

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{bmatrix}, \ \mathbf{W} = \begin{bmatrix} \mathbf{W}_1 \\ \vdots \\ \mathbf{W}_n \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_n \end{bmatrix}, \ \mathbf{e} = \begin{bmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_n \end{bmatrix}.$$

These quantities are obtained by stacking the various subject-specific vectors and matrices in a vertical manner. Also, take $\mathbf{Z} = \text{diag}\{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$. Here $\mathbf{Y}$ and $\mathbf{e}$ are $N \times 1$ vectors; $\mathbf{b}$ is an $n(1 + J + m) \times 1$ vector; $\mathbf{X}$ is an $N \times J(p + 1)$ matrix; $\mathbf{W}$ is an $N \times JQ$ matrix; and $\mathbf{Z}$ is an $N \times n(1 + J + m)$ block diagonal matrix. We can now write the joint model as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{u} + \mathbf{Z}\mathbf{b} + \mathbf{e}. \tag{A.3}$$

The assumptions for (A.1) imply that $\mathbf{Y} \sim \mathcal{N}_N(\mathbf{X}\boldsymbol{\beta}, \text{var}(\mathbf{Y}))$ with

$$\text{var}(\mathbf{Y}) = \mathbf{W}\mathbf{G}_\mathbf{u}\mathbf{W}^T + \mathbf{Z}\,\text{diag}\{\mathbf{G}_{\mathbf{b}_1}, \ldots, \mathbf{G}_{\mathbf{b}_n}\}\mathbf{Z}^T + \text{diag}\{\mathbf{R}_1, \ldots, \mathbf{R}_n\}. \tag{A.4}$$

The BLUPs of the random effects [27] in the model are

$$\begin{aligned} &E(\mathbf{u}|\mathbf{Y}) = \mathbf{G}_\mathbf{u}\mathbf{W}^T\text{var}(\mathbf{Y})^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \\ &E(\mathbf{b}_i|\mathbf{Y}) = E(\mathbf{b}_i|\mathbf{Y}_i) = \mathbf{G}_{\mathbf{b}_i}\mathbf{Z}_i^T\text{var}(\mathbf{Y}_i)^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}),\ i = 1, \ldots, n, \end{aligned} \tag{A.5}$$

where the covariance matrices involved are given by (A.2) and (A.4). The fitted values become

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{W}\hat{\mathbf{u}} + \mathbf{Z}\hat{\mathbf{b}},$$

where $\hat{\mathbf{u}}$ and $\hat{\mathbf{b}}$ are estimated BLUPs [22, Chapter 2] obtained by replacing $\boldsymbol{\theta}$ in (A.5) with $\hat{\boldsymbol{\theta}}$.

Thus far in this section the model (1) was assumed for the data, but often its special cases are needed. For example, when the model does not include the interaction term $b_{ij}$, we can represent it in the form (A.3) by respectively redefining $\mathbf{b}_i$, $\mathbf{G}_{\mathbf{b}_i}$ and $\mathbf{Z}_i$ by dropping $b_{i1}, \ldots, b_{iJ}$ from $\mathbf{b}_i$, their variances $\psi^2$ from $\mathbf{G}_{\mathbf{b}_i}$ and the corresponding columns from $\mathbf{Z}_i$; and replacing $1 + J + m$ with $1 + m$. Likewise, when the within-subject errors are independent, the matrix form of the model is same as (A.3) except that the error covariance matrix $\mathbf{R}_{ij}$ is now a diagonal matrix, obtained by setting $\rho = 0$ in its off-diagonal elements.

| | | autocorrelation model | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | parameter | IND | AR(0.2) | AR(0.5) | MA(0.2) | MA(0.5) | CS(0.2) | CS(0.5) |
| 30 | $\sigma^2_{e1}/\sigma^2_{e2}$ | 95.4 | 96.2 | 96.2 | 96.4 | 95.6 | 96.2 | 96.2 |
| | $\mu_1(t)$ | 95.0 | 94.6 | 94.6 | 93.0 | 94.4 | 93.2 | 92.6 |
| | $\mu_2(t)$ | 93.2 | 95.2 | 94.2 | 94.0 | 92.6 | 92.4 | 94.2 |
| | $\mu_2(t) - \mu_1(t)$ | 96.0 | 94.2 | 93.4 | 93.0 | 94.4 | 93.2 | 94.6 |
| | TDI$(0.9, t)$ | 95.6 | 95.6 | 93.8 | 94.2 | 94.4 | 93.6 | 93.4 |
| | CCC$(t)$ | 98.2 | 96.8 | 98.2 | 97.0 | 97.4 | 96.6 | 97.6 |
| 60 | $\sigma^2_{e1}/\sigma^2_{e2}$ | 96.2 | 95.0 | 95.4 | 93.6 | 94.8 | 95.6 | 94.8 |
| | $\mu_1(t)$ | 94.0 | 95.4 | 95.2 | 93.6 | 95.8 | 95.0 | 94.4 |
| | $\mu_2(t)$ | 95.8 | 95.4 | 94.8 | 93.8 | 91.8 | 94.2 | 94.6 |
| | $\mu_2(t) - \mu_1(t)$ | 94.0 | 92.8 | 93.4 | 94.8 | 94.8 | 94.6 | 94.4 |
| | TDI$(0.9, t)$ | 93.6 | 94.8 | 93.6 | 94.6 | 95.8 | 96.4 | 91.0 |
| | CCC$(t)$ | 96.4 | 95.8 | 98.2 | 98.4 | 97.8 | 98.4 | 96.4 |
| 100 | $\sigma^2_{e1}/\sigma^2_{e2}$ | 95.8 | 94.2 | 95.8 | 96.0 | 94.0 | 94.6 | 95.8 |
| | $\mu_1(t)$ | 93.0 | 95.0 | 95.2 | 93.0 | 95.0 | 94.2 | 94.6 |
| | $\mu_2(t)$ | 94.2 | 94.6 | 95.0 | 93.4 | 94.8 | 94.2 | 94.4 |
| | $\mu_2(t) - \mu_1(t)$ | 95.8 | 95.2 | 94.4 | 94.4 | 95.4 | 93.6 | 95.6 |
| | TDI$(0.9, t)$ | 95.4 | 95.0 | 96.0 | 95.0 | 95.8 | 94.2 | 94.8 |
| | CCC$(t)$ | 97.2 | 96.8 | 98.0 | 97.4 | 97.8 | 95.8 | 96.8 |

Table 1: Estimated individual and simultaneous coverage probabilities (in %) for 95% confidence intervals and bands in the *balanced design* case.

| | | autocorrelation model | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | parameter | IND | AR(0.2) | AR(0.5) | MA(0.2) | MA(0.5) | CS(0.2) | CS(0.5) |
| 30 | $\sigma^2_{e1}/\sigma^2_{e2}$ | 97.6 | 96.2 | 96.8 | 95.2 | 95.6 | 97.4 | 97.4 |
| | $\mu_1(t)$ | 93.8 | 93.6 | 92.0 | 94.0 | 91.4 | 91.8 | 95.6 |
| | $\mu_2(t)$ | 93.4 | 94.4 | 92.8 | 93.6 | 92.8 | 94.8 | 95.6 |
| | $\mu_2(t) - \mu_1(t)$ | 91.8 | 91.0 | 93.8 | 93.6 | 93.0 | 94.6 | 94.0 |
| | TDI$(0.9, t)$ | 94.4 | 94.6 | 94.0 | 92.6 | 94.0 | 93.4 | 94.8 |
| | CCC$(t)$ | 97.8 | 98.8 | 98.2 | 98.2 | 97.8 | 97.4 | 98.2 |
| 60 | $\sigma^2_{e1}/\sigma^2_{e2}$ | 95.2 | 94.2 | 95.2 | 96.8 | 95.0 | 94.6 | 94.8 |
| | $\mu_1(t)$ | 94.8 | 95.0 | 94.8 | 95.2 | 96.2 | 94.8 | 94.8 |
| | $\mu_2(t)$ | 95.6 | 94.6 | 94.2 | 92.4 | 95.4 | 95.6 | 94.2 |
| | $\mu_2(t) - \mu_1(t)$ | 94.6 | 95.2 | 92.8 | 91.6 | 94.4 | 93.2 | 94.0 |
| | TDI$(0.9, t)$ | 95.2 | 95.0 | 94.2 | 94.2 | 95.4 | 93.6 | 94.4 |
| | CCC$(t)$ | 98.0 | 96.8 | 98.0 | 97.2 | 98.8 | 98.4 | 98.0 |
| 100 | $\sigma^2_{e1}/\sigma^2_{e2}$ | 95.8 | 93.8 | 95.0 | 94.8 | 94.2 | 94.0 | 95.6 |
| | $\mu_1(t)$ | 93.2 | 94.8 | 94.4 | 91.2 | 95.2 | 93.8 | 95.8 |
| | $\mu_2(t)$ | 94.4 | 94.8 | 94.2 | 93.2 | 94.4 | 96.0 | 94.8 |
| | $\mu_2(t) - \mu_1(t)$ | 94.8 | 95.0 | 96.4 | 92.6 | 94.6 | 94.4 | 94.6 |
| | TDI$(0.9, t)$ | 95.2 | 95.4 | 96.2 | 95.2 | 96.8 | 95.4 | 94.8 |
| | CCC$(t)$ | 97.2 | 96.6 | 97.4 | 96.4 | 97.8 | 96.8 | 97.0 |

Table 2: Estimated individual and simultaneous coverage probabilities (in %) for 95% confidence intervals and bands in the *unbalanced design* case.

| parameter | estimate | 95% interval |
|---|---|---|
| $\sigma_b$ | 3.05 | (2.60, 3.59) |
| $\eta$ | 0.69 | (0.59, 0.82) |
| $\sigma_{e1}$ | 2.33 | (2.12, 2.57) |
| $\sigma_{e2}$ | 2.06 | (1.87, 2.27) |
| $\rho$ | 0.55 | (0.49, 0.62) |

Table 3: Estimates and 95% confidence intervals for variance-covariance related parameters for body fat data.

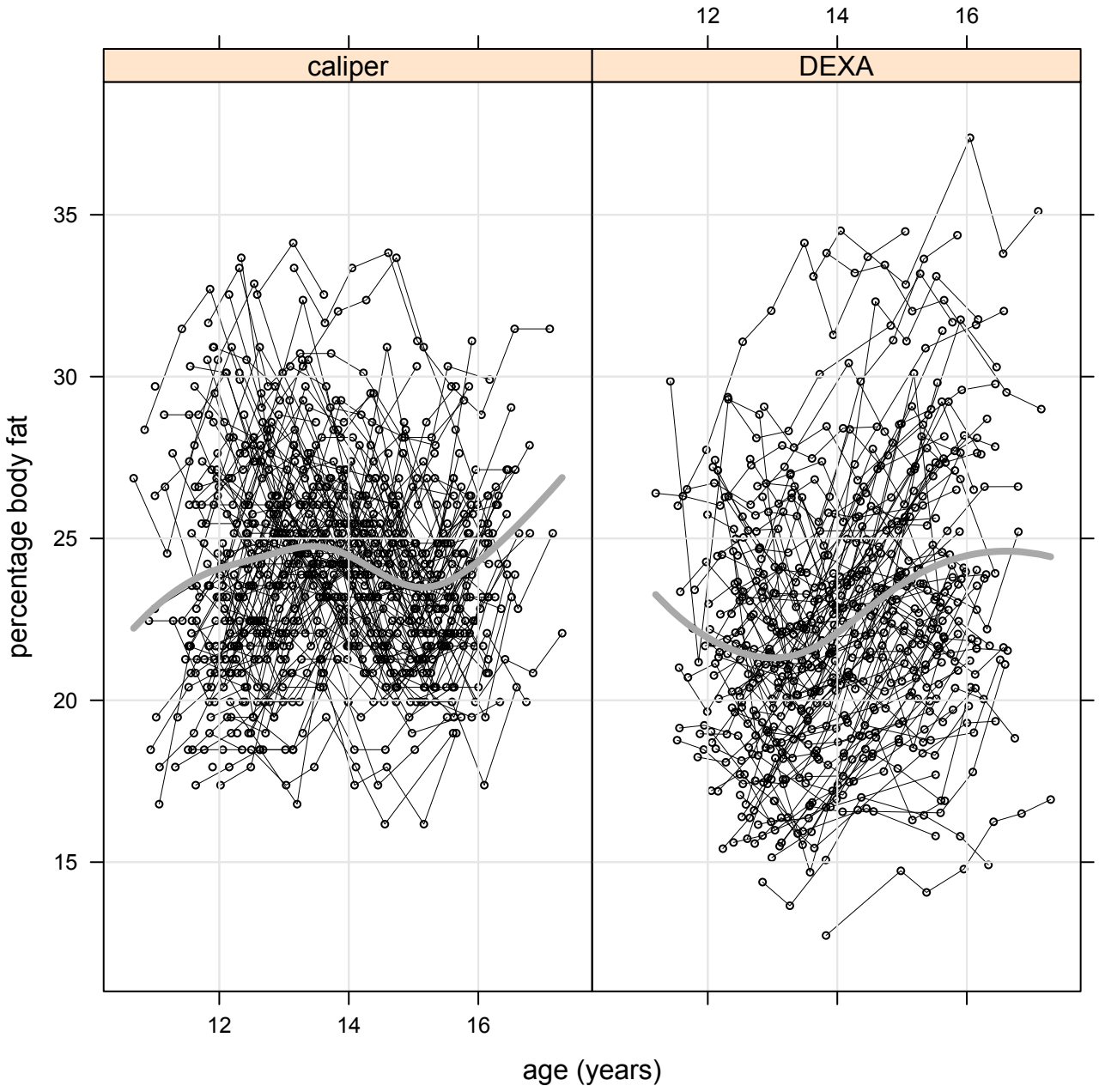


Figure 1: Trajectories of percentage body fat measurements from caliper and DEXA methods. The points on the same trajectory are connected by lines. The thick grey curves in the middle are the estimated mean functions.

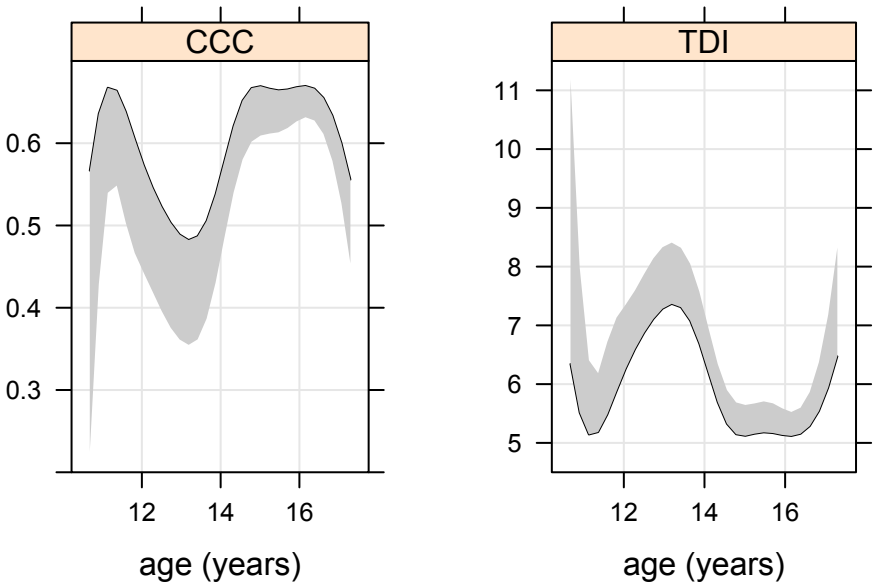


Figure 2: Left panel: Estimate of CCC (solid curve) and its 95% simultaneous lower confidence band (shaded region). Right panel: Estimate of TDI with $p_0 = 0.90$ (solid curve) and its 95% simultaneous upper confidence band (shaded region).

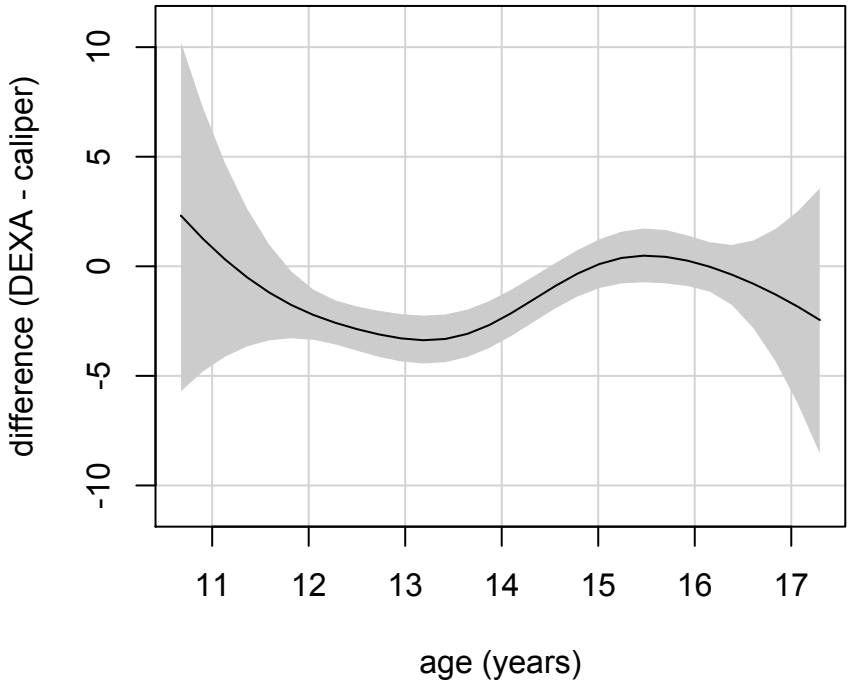


Figure 3: Estimate of difference in means of DEXA and caliper methods (solid curve) and its 95% simultaneous two-sided confidence band (shaded region).