Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 138 (2008) 1102–1115
journal of statistical planning and inference
www.elsevier.com/locate/jspi

# A tolerance interval approach for assessment of agreement in method comparison studies with repeated measurements

Pankaj K. Choudhary*

*Department of Mathematical Sciences, University of Texas at Dallas, EC 35, P.O. Box 830688, Richardson, TX 75083-0688, USA*

Received 23 December 2005; received in revised form 31 July 2006; accepted 28 March 2007
Available online 18 May 2007

**Abstract**

This paper generalizes the tolerance interval approach for assessing agreement between two methods of continuous measurement for repeated measurement data—a common scenario in applications. The repeated measurements may be longitudinal or they may be replicates of the same underlying measurement. Our approach is to first model the data using a mixed model and then construct a relevant asymptotic tolerance interval (or band) for the distribution of appropriately defined differences. We present the methodology in the general context of a mixed model that can incorporate covariates, heteroscedasticity and serial correlation in the errors. Simulation for the no-covariate case shows good small-sample performance of the proposed methodology. For the longitudinal data, we also describe an extension for the case when the observed time profiles are modelled nonparametrically through penalized splines. Two real data applications are presented.


*Keywords:* Concordance correlation; Limits of agreement; Method comparison; Mixed model; Penalized splines; Tolerance interval; Total deviation index

## 1. Introduction

We consider the problem of assessment of agreement between two methods—a reference method $y_1$ and a test method $y_2$, for measuring a continuous response variable when there are repeated measurements from the methods. This problem arises in medical applications where $y_1$ is generally expensive or invasive, and $y_2$ is a more convenient alternative than $y_1$. The main goal of method comparison is to determine whether the measurements from $y_1$ and $y_2$ on an individual can be interchanged without leading to any inconsistency in the interpretation of the measured response. The repeated measurements may be longitudinal or may be replicates of the same underlying measurement. In the latter case, we can also assess the agreement of a method to itself. This intra-method agreement is also known as the *repeatability* of a method. The evaluation of intra-method agreement is important because if the methods do not agree well with themselves, they cannot be expected to agree well with each other (see, e.g., Bland and Altman, 1999;

* Tel.: +1 972 883 4436; fax: +1 972 883 6622.
*E-mail address:* pankaj@utdallas.edu.

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.03.056

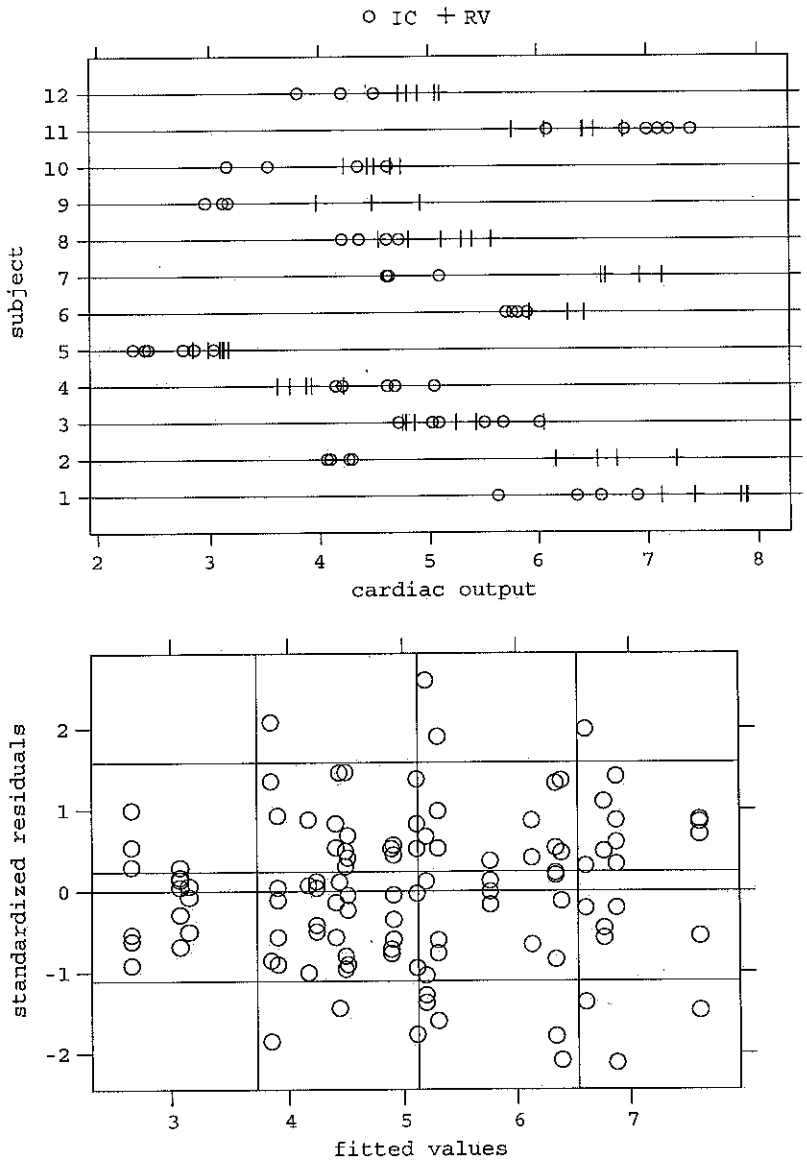


Fig. 1. Top: Plot of cardiac output measurements on each individual from RV and IC methods. Bottom: Residual plot of the fitted model (see Section 5.1).

Hawkins, 2002). The extent of intra-method agreement serves as a benchmark for assessing the agreement between methods. We now describe two real examples that motivated this work.

### *1.1. Cardiac output data*

In this example from Bland and Altman (1999), cardiac output is measured in 12 individuals from two methods—radionuclitide ventriculography (RV, $y_1$) and impedance cardiography (IC, $y_2$)—with the goal of assessing their agreement. Both methods have equal number of replicate measurements on an individual, but this number varies between 3 and 6. There is a total of 120 measurements (60 from each method). Fig. 1 plots these data. We get the impression that IC measurements tend to be smaller and have a larger within-individual variation than their RV counterparts. There is also a strong evidence for method-individual interaction. Both methods seem to have good repeatability but their agreement does not appear strong. We discuss the inference in Section 5.1 based on a mixed model fitted to these data.

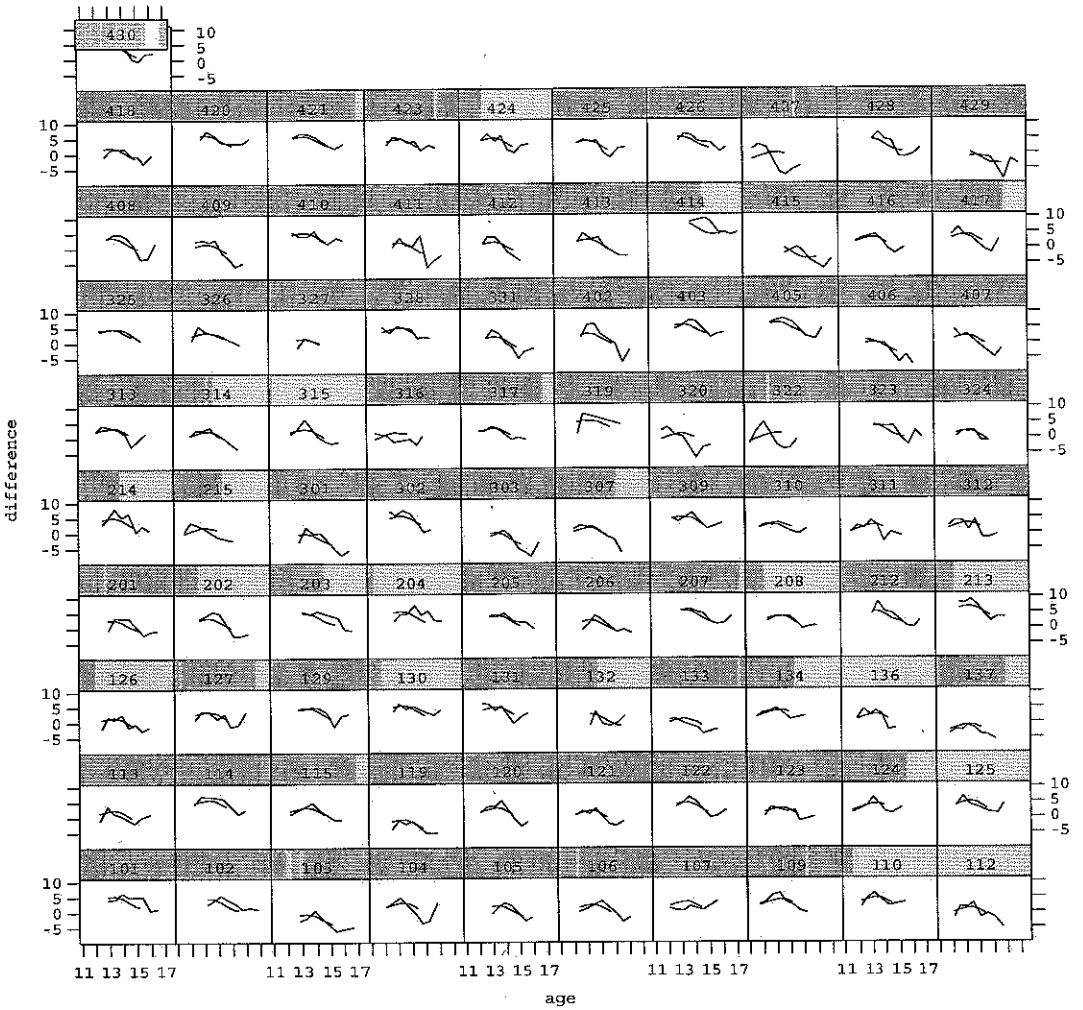


Fig. 2. Observed (solid curve) versus fitted (broken curve) time profiles of the differences ($y_1 - y_2$) of the percent body fat measurements from skinfold calipers ($y_1$) and DEXA ($y_2$). Section 5.2 describes the fitted semiparametric model. The numbers 101–430 represent the individual ID's.

### *1.2. Body fat data*

In this Young Women's Health Study example from Chinchilli et al. (1996), percent body fat is measured over time using two methods—skinfold calipers ($y_1$) and dual energy X-ray absorptiometry (DEXA, $y_2$)—on a cohort of 91 adolescent girls. Their initial visit occurred around age 12, and there were 8 subsequent visits roughly 6 months apart. The measurements here are longitudinal and are paired over time. There are 657 complete pairs available, but for the illustration, we consider only 654 pairs after excluding 3 outliers. The excluded observations had unusually large normalized residuals for the model described in Section 5.2, which fits quite well to the remaining data. We have between 4 and 8 repeated measurements on each girl. The observed percent body fat varies between 12% and 38%. Our goal is to estimate the extent of agreement between the two methods as a function of age. For simplicity, we will focus the modelling effort on the differences ($y_1 - y_2$) of the paired measurements as a function of the covariate age at the time of visit. Fig. 2 shows their observed time profiles. In general, the differences do not appear small compared to the magnitude of the body fat measurements. These profiles have non-linear features that are hard to model using a low degree polynomial. So we model them nonparametrically using penalized splines via their mixed model representation. Section 5.2 provides the details of the fitted semiparametric model and the resulting agreement evaluation.

For assessing agreement, we focus on the tolerance interval methodology of Lin (2000), Lin et al. (2002) and Choudhary and Nagaraja (2007). In the i.i.d. case, the observed data consist of $m$ pairs of measurements from $(y_1, y_2)$, say, $(y_{i1}, y_{i2})$, $i = 1, \ldots, m$. The analysis focuses on the differences $d_i = y_{i1} - y_{i2}$, $i = 1, \ldots, m$, which are assumed to be a random sample from the difference population $d = y_1 - y_2$ that follows a normal distribution. The $p_0$th percentile of $|d|$, say $q$, is taken as the measure of agreement between the methods. Lin (2000) introduced this measure as the *total deviation index*. Smaller values of $q$ indicate better agreement. Here $p_0 (> 0.5)$ is a large probability cutoff specified by the practitioner. For inference, we construct a level $(1 - \alpha)$ upper confidence bound (UCB) $U$ for the parameter $q$. The interval $[-U, U]$ then becomes a $p_0$ probability content *tolerance interval* for the distribution of $d$, i.e., we have

$$\Pr\{F(U) - F(-U) \geqslant p_0\} = 1 - \alpha,$$

with $F(\cdot)$ as the cumulative distribution function of $d$. This tolerance interval essentially estimates the range of $p_0$ proportion of the population of measurement differences. When it does not contain any large clinically meaningful differences, the practitioner infers sufficient agreement between the methods.

Choudhary and Ng (2006) generalized this basic methodology for the normal theory regression setup where the mean or the variance of $d$ depends on a known, continuous covariate $x \in \mathfrak{X}$ considered as non-random. This covariate is generally the observed average measurement that serves as a proxy for the magnitude of the true unobservable measurement. Let $d_x$ denote the population of $y_1 - y_2$ differences at $x$. The agreement at $x$ is measured by the $p_0$th percentile of $|d_x|$, say $q_x$. The authors develop an asymptotic UCB $U_x$ for $q_x$ that has simultaneous confidence $(1 - \alpha)$ over $\mathfrak{X}$. For the band $[-U_x, U_x]$, $x \in \mathfrak{X}$, we now have

$$\Pr\{F_x(U_x) - F_x(-U_x) \geqslant p_0, \text{ for all } x \in \mathfrak{X}\} \approx 1 - \alpha, \tag{1}$$

where $F_x(\cdot)$ is the cumulative distribution function of $d_x$. This band can be interpreted as a $p_0$-content tolerance band with simultaneous confidence $(1 - \alpha)$. It estimates the extent of $p_0$ proportion of differences in measurements from the two methods, which now depends on $x$. The practitioner uses this band in the same way as the tolerance interval for the identical distribution case, but the inference now is simultaneously valid over entire $\mathfrak{X}$.

The goal of this paper is to further generalize this methodology to incorporate repeated measurements from the methods. Let $y_{ijk}$ denote the $k$th measurement from the $j$th method on the $i$th individual ($i = 1, \ldots, m$; $j = 1, 2$; $k = 1, \ldots, n_{ij}$). Here the data may be unbalanced, i.e., $n_{i1}$ and $n_{i2}$ may be unequal; and the measurements need not be paired. In other words, there is no requirement that the measurements from the two methods be collected together or at the same time. Moreover, the distributions of measurements may depend on one or more covariates. We also discuss a tolerance interval type measure for the assessment of intra-method agreement in Section 3. Although, measures such as the intra-class correlation (see, e.g., Fleiss, 1986, Chapter 1) are also appropriate for this purpose, the main advantage of a tolerance interval type measure is that it serves as a benchmark against which one can evaluate the inter-method agreement using a tolerance interval.

Our strategy is to first model the observed data using a mixed model as described in Section 2. The mixed models provide a popular framework for modelling the repeated measurement data (see, e.g., Diggle et al., 2002, Chapter 9). Once we have a model for the data, we proceed to the assessment of agreement in Section 3. Here a key issue is to define an appropriate difference population whose distribution contains the information regarding agreement. When the data are not paired, a direct difference between measurements from the two methods is not well defined. We address this issue by using the assumed model for the data to define the difference population. Section 4 contains a simulation study of the proposed methodology. In Section 5, we revisit the above examples to illustrate the application. We conclude in Section 6 with a discussion.

We will use the bold-face notation for vectors and matrices. All vectors are column vectors unless noted otherwise. The transpose of matrix $\mathbf{X}$ will be denoted as $\mathbf{X}'$. We will use $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and non-singular covariance matrix $\boldsymbol{\Sigma}$. The dimension of the distribution will be clear from the context. All the computations and the data analysis reported in this paper have been performed using the statistical software R (R Development Core Team, 2004). We have used the `nlme` package of Pinheiro and Bates (2000) in R for fitting mixed models.

## 2. Modelling the observed data

### 2.1. General case

Let $\mathbf{y}_{ij}$ be the column vector of $n_{ij}$ measurements on the $i$th individual from the $j$th method. We assume a linear mixed model for $\mathbf{y}_{ij}$ of the following general form:

$$\mathbf{y}_{ij} = \mathbf{P}_{ij}\boldsymbol{\phi} + \mathbf{W}_{ij}\mathbf{v}_i + \mathbf{X}_{ij}\boldsymbol{\beta}_j + \mathbf{Z}_{ij}\mathbf{b}_{ij} + \boldsymbol{\epsilon}_{ij}; \quad j = 1, 2, \ i = 1, \ldots, m. \tag{2}$$

Here $\mathbf{P}_{ij}$, $\mathbf{W}_{ij}$, $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ are full-rank design matrices; and $\boldsymbol{\epsilon}_{ij}$ is the column vector of errors. Further, $\boldsymbol{\phi}$, $\boldsymbol{\beta}_j$, $\mathbf{v}_i$ and $\mathbf{b}_{ij}$ are also column vectors—respectively representing the fixed-effects common to both methods, the fixed-effects specific to the $j$th method, the random-effects of $i$th individual common to both methods and the random-effects of $i$th individual specific to the $j$th method. The design matrices are such that a fixed intercept for each method and a random intercept for each individual are included. We assume that

$$\mathbf{v}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \quad \begin{bmatrix}\mathbf{b}_{i1}\\ \mathbf{b}_{i2}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Psi}_{11} & \boldsymbol{\Psi}_{12}\\ \boldsymbol{\Psi}_{12} & \boldsymbol{\Psi}_{22}\end{bmatrix}\right), \quad \begin{bmatrix}\boldsymbol{\epsilon}_{i1}\\ \boldsymbol{\epsilon}_{i2}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Lambda}_{i1} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Lambda}_{i2}\end{bmatrix}\right).$$

These vectors are independent for different $i$ and are also mutually independent. All the covariance matrices are positive-definite. They are generally defined in terms of a small number of parameters. It follows that the joint distribution of $(\mathbf{y}_{i1}, \mathbf{y}_{i2})$, after integrating out the individual random effects, is multivariate normal with $E(\mathbf{y}_{ij}) = \mathbf{P}_{ij}\,\boldsymbol{\phi} + \mathbf{X}_{ij}\boldsymbol{\beta}_j$, $var(\mathbf{y}_{ij}) = \mathbf{W}_{ij}\boldsymbol{\Sigma}\mathbf{W}'_{ij} + \mathbf{Z}_{ij}\boldsymbol{\Psi}_{jj}\mathbf{Z}'_{ij} + \boldsymbol{\Lambda}_{ij}$ and $cov(\mathbf{y}_{i1}, \mathbf{y}_{i2}) = \mathbf{W}_{i1}\boldsymbol{\Sigma}\,\mathbf{W}'_{i2} + \mathbf{Z}_{i1}\boldsymbol{\Psi}_{12}\mathbf{Z}'_{i2}$. The distributions are independent for different $i$. Additionally, to make the methods comparable, we must also assume that the marginal models for $y_1$ and $y_2$ are similar in that their mean vectors and covariance matrices are parameterized identically. The two marginal models may differ in the values of parameters, but the parameters have the same interpretation for the two models. Thus, in particular, the columns of the design matrices correspond to identical effects. Hence the quantities

$$\boldsymbol{\beta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2, \quad \mathbf{b}_i = \mathbf{b}_{i1} - \mathbf{b}_{i2}, \quad var(\mathbf{b}_i) = \boldsymbol{\Psi} = \boldsymbol{\Psi}_{11} + \boldsymbol{\Psi}_{12} - 2\boldsymbol{\Psi}_{12} \tag{3}$$

are meaningful, and $\mathbf{b}_i \sim$ independent $\mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$. It is, however, not required that the design matrices be the same for each $j$. For example, they may have different number of rows.

In the model (2), we have made a distinction between two types of effects. The first are $\boldsymbol{\phi}$ and $\mathbf{v}_i$ that are common to both methods. They may represent the effects of multiple covariates. The second are $\boldsymbol{\beta}_j$ and $\mathbf{b}_{ij}$ that vary with method. The design matrices $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ for these effects, and in addition, the error covariance matrix $\boldsymbol{\Lambda}_{ij}$ in (2), may also involve a continuous covariate $x$. We call it *agreement covariate* to distinguish it from the covariates in $\mathbf{P}_{ij}$ and $\mathbf{W}_{ij}$. We will see in Section 3 that it plays a role in defining the difference population that we focus on for agreement assessment. In contrast, the effects of other covariates cancel out. For simplicity, we assume that there is at most one agreement covariate. Extension to multiple agreement covariates is possible, but its details get rather messy and is typically not needed in practice. We additionally assume that $\boldsymbol{\Lambda}_{ij}$ depends on $i$ only through the value of $x$. Thus $\boldsymbol{\Lambda}_{ij}$ actually denotes $\boldsymbol{\Lambda}_{x_i j}$. In applications, this $x$ is frequently a proxy for the true magnitude of measurement or the time of measurement.

### 2.2. Paired measurements case

When the measurements $(y_{i1k}, y_{i2k})$ are collected together in pairs for each $k$, the difference $d_{ik} = y_{i1k} - y_{i2k}$ is well defined. Here in addition to (3), we also have $n_{i1} = n_{i2} = n_i$ (say), $\mathbf{X}_{i1} = \mathbf{X}_{i2} = \mathbf{X}_i$ (say), $\mathbf{Z}_{i1} = \mathbf{Z}_{i2} = \mathbf{Z}_i$ (say), and the other design matrices are also the same for each $j$. Furthermore, the differences

$$\mathbf{d}_i = \mathbf{y}_{i1} - \mathbf{y}_{i2}, \quad \boldsymbol{\epsilon}_i = \boldsymbol{\epsilon}_{i1} - \boldsymbol{\epsilon}_{i2}, \quad i = 1, \ldots, m \tag{4}$$

are also meaningful, and $\boldsymbol{\epsilon}_i \sim$ independent $\mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_i = \boldsymbol{\Lambda}_{i1} + \boldsymbol{\Lambda}_{i2})$. In this paired case, we can focus on modelling the differences directly rather than modelling the individual measurements since our ultimate goal is to use the population of differences to evaluate the agreement between the two methods. To this end, let

$$\mathbf{d}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, m, \tag{5}$$

where $\boldsymbol{\beta}$, $\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i$ are defined by (3)–(4). One can think of this model as $\mathbf{d}_i = \mathbf{y}_{i1} - \mathbf{y}_{i2}$ with $\mathbf{y}_{ij}$ given by (2). However, the estimates of parameters in (5) may not be the same as those obtained by fitting (2). The distributional assumptions

for (2) also give us, $\mathbf{d}_i \sim$ independent $\mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i' + \boldsymbol{\Lambda}_i)$, $i = 1, \ldots, m$. As in (2), an agreement covariate $x$ may be involved in $\mathbf{X}_i$, $\mathbf{Z}_i$ or $\boldsymbol{\Lambda}_i$. The effects $\boldsymbol{\phi}$ and $\mathbf{v}_i$ do not appear in (5) as they cancel out upon differencing.

### 2.3. Parameter estimation

Whether we use the model (2) for individual measurements or the model (5) for differences, in either case, let $\boldsymbol{\theta}$ be the vector of all the model parameters. We will use the method of maximum likelihood to fit the desired model and denote the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ as $\hat{\boldsymbol{\theta}}$. See Pinheiro and Bates (2000, Chapters 2, 5) for an excellent account of the computational details involved in evaluating the likelihood function of a mixed model and its maximization. They also discuss the issue of how to parameterize the random-effect covariance matrix and give various examples of the variance and covariance structures that can be incorporated in the error covariance matrix. We do not consider the restricted maximum likelihood estimation since it does not lead to a joint distribution of the estimates of fixed-effects and variance–covariance parameters, which we need for the assessment of agreement. Finally, let $\mathbf{I}$ denote the observed Fisher information matrix—the matrix of second order partial derivatives of the negative log-likelihood function of the desired model (2) or (5) with respect to $\boldsymbol{\theta}$ and evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. A closed-form expression for $\mathbf{I}$ is not available in general but it can be easily computed numerically. When $m$ is large, it is well known that $\hat{\boldsymbol{\theta}}$ approximately follows $\mathcal{N}(\boldsymbol{\theta}, \mathbf{I}^{-1})$.

## 3. Assessment of agreement

### 3.1. Agreement between methods

We now describe how to assess the agreement between $y_1$ and $y_2$ assuming that the data are modelled using either (2) or (5). First we deal with the case when an agreement covariate $x \in \mathfrak{X}$ is included in the model either for modelling the mean part or the variance part of the data. In applications, $\mathfrak{X}$ is generally the range of observed values of $x$.

Let the row vectors $\mathbf{X}_x$ and $\mathbf{Z}_x$ respectively denote a general row of the design matrices $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ corresponding to the covariate value $x$. Similarly, let the row vectors $\mathbf{P}$ and $\mathbf{W}$ respectively represent a general row of the design matrices $\mathbf{P}_{ij}$ and $\mathbf{W}_{ij}$ corresponding to any fixed setting of the covariates involved. Next, let $(y_{x1}, y_{x2})$ denote the bivariate population of $(y_1, y_2)$ measurements at these covariate settings. The dependence of $(y_{x1}, y_{x2})$ on covariates other than $x$ is suppressed for notational convenience. From our assumptions for (2), it follows that $(y_{x1}, y_{x2})$ has a bivariate normal distribution with $E(y_{xj}) = \mathbf{P}\boldsymbol{\phi} + \mathbf{X}_x\boldsymbol{\beta}_j$, $var(y_{xj}) = \mathbf{W}\boldsymbol{\Sigma}\mathbf{W}' + \mathbf{Z}_x\boldsymbol{\Psi}_{jj}\mathbf{Z}_x' + \Lambda_{xj}$, $j = 1, 2$, and $cov(y_{x1}, y_{x2}) = \mathbf{W}\boldsymbol{\Sigma}\mathbf{W}' + \mathbf{Z}_x\boldsymbol{\Psi}_{12}\mathbf{Z}_x'$. Here $\Lambda_{xj}$ is the error variance of method $j$ evaluated at $x$. This joint distribution is defined irrespective of whether the data are balanced or not and whether the measurements are paired or not. Now, let $d_x = y_{x1} - y_{x2}$ denote the population of differences. It follows that

$$d_x \sim \mathcal{N}(\mu_x = \mathbf{X}_x\boldsymbol{\beta}, \sigma_x^2 = \mathbf{Z}_x\boldsymbol{\Psi}\mathbf{Z}_x' + \Lambda_x), \quad x \in \mathfrak{X}, \tag{6}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\Psi}$ are defined in (3), and $\Lambda_x = \Lambda_{x1} + \Lambda_{x2}$. This distribution does not involve the effects $\boldsymbol{\phi}$ and $\mathbf{v}_i$ in (2) that are common to both methods. It contains the information regarding the agreement between $(y_1, y_2)$, and for the tolerance interval approach, we focus on it for inference. In particular, we take the $p_0$th percentile function of $|d_x|$, defined as

$$q_x = \sigma_x\{\chi_1^2(p_0, \mu_x^2/\sigma_x^2)\}^{1/2}, \quad x \in \mathfrak{X} \tag{7}$$

as the measure of agreement at $x$. Here the large probability $p_0$ is assumed to be specified in advance by the practitioner; $(\mu_x, \sigma_x^2)$ are defined in (6); and $\chi_1^2(p_0, \Delta)$ represents the $p_0$th percentile of a $\chi^2$-distribution with one degree of freedom and non-centrality parameter $\Delta$.

Once an appropriate parsimonious model of the form (2) or (5) is fitted to the data and the MLE $\hat{\boldsymbol{\theta}}$ is available, the MLE $\hat{q}_x$ of $q_x$ is simply obtained by substitution, i.e.,

$$\hat{q}_x = \hat{\sigma}_x\{\chi_1^2(p_0, \hat{\mu}_x^2/\hat{\sigma}_x^2)\}^{1/2}, \quad x \in \mathfrak{X}$$

with $(\hat{\mu}_x = \mathbf{X}_x\hat{\boldsymbol{\beta}}, \hat{\sigma}_x^2 = \mathbf{Z}_x\hat{\boldsymbol{\Psi}}\mathbf{Z}_x' + \hat{\Lambda}_x)$ as the MLEs of $(\mu_x, \sigma_x^2)$. To derive a simultaneous UCB of $q_x$ for $x \in \mathfrak{X}$ so that (1) holds, we consider the curve

$$U_x = \exp\{\log \hat{q}_x - c_\alpha(\mathbf{G}_x'\mathbf{I}^{-1}\mathbf{G}_x)^{1/2}\}, \quad x \in \mathfrak{X}, \tag{8}$$

where $\mathbf{G}_x$ is the vector of partial derivatives of $\log q_x$ with respect to $\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$; and $c_\alpha$ $(< 0)$ is the critical point so that the large-sample simultaneous confidence level of this UCB is $(1 - \alpha)$. Recall that $\mathbf{I}$ here is the observed information matrix for the fitted model. The motivation for this curve comes from the realization that $\hat{q}_x$, being an MLE, is approximately normal and, for small samples, this approximation is more accurate on the $\log q_x$ scale. The quantity $\mathbf{G}_x'\mathbf{I}^{-1}\mathbf{G}_x$ estimates the asymptotic variance of $\log \hat{q}_x$. To obtain the critical point, we follow Choudhary and Ng (2006) and solve the equation

$$\alpha = \Pr(t_\nu \leqslant c_\alpha) + \frac{\kappa_0}{2\pi}\left(1 + \frac{c_\alpha^2}{\nu}\right)^{-\nu/2}, \quad \nu = (m - l) \tag{9}$$

for $c_\alpha$. Here the random variable $t_\nu$ follows a $t$-distribution with $\nu$ degrees of freedom, $l$ is the dimension of $\boldsymbol{\beta}$, and $\kappa_0$ is defined as

$$\kappa_0 = \int_{\mathfrak{X}} \frac{1}{\mathbf{L}_x'\mathbf{L}_x}((\mathbf{L}_x'\mathbf{L}_x)(\dot{\mathbf{L}}_x'\dot{\mathbf{L}}_x) - (\mathbf{L}_x'\dot{\mathbf{L}}_x)^2)^{1/2}\,\mathrm{d}x$$

by taking $\mathbf{L}_x = \mathbf{I}^{-1/2}\mathbf{G}_x$ and $\dot{\mathbf{L}}_x = \mathbf{I}^{-1/2}(\partial/\partial x)\mathbf{G}_x$, with the partial differentiation applied elementwise. This critical point is easy to compute numerically. A more accurate alternative for small samples is to use a parametric bootstrap-$t$ method (see, e.g., Davison and Hinkley, 1997, Chapter 5) for computing the critical point. It involves the following steps:

1. Simulate $m$ independent draws from the model fitted to the original data using $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and the observed covariate values. Denote them as $(\mathbf{y}_{11}^*, \mathbf{y}_{12}^*), \ldots, (\mathbf{y}_{m1}^*, \mathbf{y}_{m2}^*)$ if the model (2) is fitted or as $\mathbf{d}_1^*, \ldots, \mathbf{d}_m^*$ if the model (5) is fitted. This sample is a parametric resample of the original data.
2. Fit the relevant model (2) or (5) to the resample in Step 1, and compute the MLE of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}^*$, the associated observed information matrix, say $\mathbf{I}^*$, the gradient vector, say $\mathbf{G}_x^*$, and an estimate of $\inf_{x\in\mathfrak{X}}(\log \hat{q}_x - \log q_x)/(\mathbf{G}_x'\mathbf{I}^{-1}\mathbf{G}_x)^{1/2}$, say

$$M = \min_{1\leqslant i\leqslant m}(\log \hat{q}_{x_i}^* - \log \hat{q}_{x_i})/(\mathbf{G}_{x_i}^{*\prime}\mathbf{I}^{*-1}\mathbf{G}_{x_i}^*)^{1/2},$$

   where $x_1, \ldots, x_m$ are the observed value of $x$ in the original sample.
3. Repeat Steps 1 and 2 a large number of times, say $B$, to simulate $B$ realizations of $M$. Take the $\alpha$th sample percentile of $M$ as the critical point $c_\alpha$ in (8).

Thus far we have assumed that the model includes covariates, at least an agreement covariate $x$. An important special case results when there are no covariates (see, e.g., the model (11) in Section 4). In this case, both the vectors $\mathbf{X}_x$ and $\mathbf{Z}_x$ reduce to the scalar 1, and the variance $\Lambda_x$ is free of $x$. Consequently, the distribution of $d_x$, the population of differences between $(y_1, y_2)$ measurements, does not depend on $x$. Further from (6), $d_x \sim \mathcal{N}(\mu_x = \beta, \sigma_x^2 = \Psi + \Lambda_x)$. The agreement in this situation can be assessed by simply noting that now $(\mu_x, \sigma_x^2, q_x)$ and hence $(\hat{\mu}_x, \hat{\sigma}_x^2, \hat{q}_x)$ are constants with respect to $x$. Moreover, the question of a *simultaneous* UCB does not arise, and hence $U_x$ in (8), also a constant with respect to $x$, can be computed by using $t_{m-l}(\alpha)$, the $\alpha$th percentile of a $t_{m-l}$ distribution, as the critical point $c_\alpha$. In this case, we have a tolerance *interval* instead of a tolerance *band*. When $m$ is not large, the bootstrap-$t$ critical point is more accurate (see Section 4).

### 3.2. Agreement of a method to itself

We now consider a measure for assessing intra-method agreement whose interpretation is similar to the measure $q_x$ of inter-method agreement. Let (2) be the model for the measurements $y_{ijk}$. Assume additionally that the values of covariates on an individual do not change with $k$ and that there is no serial correlation in the within-individual

errors. These assumptions ensure that the repeated measurements on an individual are identically distributed and they provide replications of the same underlying measurement for that individual—making the assessment of repeatability possible. Let the random variable $d_{xj}$ denote the population of difference between any two replicates of $y_{xj}$ on the same individual. Under our assumptions, $d_{xj}$ is just the population difference between two uncorrelated within-individual errors associated with the two replicates. Hence, we have

$$d_{xj} \sim \mathcal{N}(\mu_{xj} = 0, \sigma_{xj}^2 = 2\Lambda_{xj}), \quad j = 1, 2,$$

where $\Lambda_{xj}$ is the within-individual error variance of the $j$th method at $x$.

In analogy with $q_x$ of (7), the $p_0$th percentile of $|d_{xj}|$, say $q_{xj} = \sigma_{xj}\{\chi_1^2(p_0, 0)\}^{1/2}$, is taken as the measure of repeatability of method $j$. This measure depends on the parameter $\theta$ of the model (2) only through $\sigma_{xj}^2$. It can be estimated as $\hat{q}_{xj} = \hat{\sigma}_{xj}\{\chi_1^2(p_0, 0)\}^{1/2}$ and its simultaneous level $(1-\alpha)$ UCB can be approximated as

$$U_{xj} = \exp\{\log \hat{q}_{xj} - c_{\alpha j}(\mathbf{G}_{xj}'\mathbf{I}^{-1}\mathbf{G}_{xj})^{1/2}\}, \quad x \in \mathfrak{X} \text{ with } \mathbf{G}_{xj} = (\partial \log q_{xj}/\partial x)_{\theta=\hat{\theta}} \tag{10}$$

with $c_{\alpha j}$ computed as in the previous section. The case when there are no covariates can also be handled on the lines of the previous section.

## 4. Monte Carlo simulation studies

In this section, we use simulation to get some insight into the small-sample coverage probabilities of the asymptotic UCB's proposed in Section 3. For simplicity, our investigation will focus only on the model,

$$y_{ijk} = \beta_j + b_{ij} + \varepsilon_{ijk}; \quad k = 1, \ldots, n_{ij}, \ j = 1, 2, \ i = 1, \ldots, m. \tag{11}$$

This model is for the situation when the repeated measurements from a method are identically distributed and there are no covariates. We will use it in Section 5.1 for the cardiac output data. Here, we assume that $(b_{i1}, b_{i2})$ follows an independent bivariate normal distribution with mean zero, variance $(\Psi_{11}, \Psi_{22})$ and covariance $\Psi_{12}$; the error $\varepsilon_{ijk}$ follows an independent $\mathcal{N}(0, \lambda_j)$ distribution; and the random-effects and the errors are mutually independent. This model can also be written as

$$\begin{bmatrix} \mathbf{y}_{i1} \\ \mathbf{y}_{i2} \end{bmatrix} \sim \text{independent } \mathcal{N}\left(\begin{bmatrix} \beta_1 \mathbf{1}_{i1} \\ \beta_2 \mathbf{1}_{i2} \end{bmatrix}, \begin{bmatrix} \Psi_{11}\mathbf{1}_{i1}\mathbf{1}_{i1}' + \lambda_1 \mathbf{J}_{i1} & \Psi_{12}\mathbf{1}_{i1}\mathbf{1}_{i2}' \\ \Psi_{12}\mathbf{1}_{i1}\mathbf{1}_{i2}' & \Psi_{22}\mathbf{1}_{i2}\mathbf{1}_{i2}' + \lambda_2 \mathbf{J}_{i2} \end{bmatrix}\right), \tag{12}$$

where $\mathbf{1}_{ij}$ is the $n_{ij}$-dimensional vector of ones and $\mathbf{J}_{ij}$ is the $n_{ij} \times n_{ij}$ identity matrix. Essentially it assumes that the measurements from the method $j$ on an individual are equi-correlated $\mathcal{N}(\beta_j, \Psi_{jj} + \lambda_j)$ random variables with intra-class correlation $\Psi_{jj}/(\Psi_{jj} + \lambda_j)$; any two measurements on the same individual but different methods have the correlation $\Psi_{12}/((\Psi_{11} + \lambda_1)(\Psi_{22} + \lambda_2))^{1/2}$; and measurements on different individuals are independent.

This model has seven parameters—$(\beta_1, \beta_2)$, $(\lambda_1, \lambda_2)$ and $(\Psi_{11}, \Psi_{12}, \Psi_{22})$. For the simulation, we take $\alpha = 0.05$, $p_0 = 0.80$, $n_{ij} \equiv n \in \{2, 3, 5\}$, $m \in \{15, 30\}$, and the parameter values $(\beta_1, \lambda_1, \Psi_{11}) = (0, 1, 16)$, $\Psi_{12} = 15.95$, $\beta_2 \in \{0, 2\}$, $\lambda_2 \in \{1, 1.5\}$ and $\Psi_{22} \in \{16, 20\}$. There is no loss of generality in taking $(\beta_1, \lambda_1) = (0, 1)$. Here $\Psi_{11} = 16 = \Psi_{22}$ corresponds to a high correlation ($> 0.90$) between methods and $(\Psi_{11}, \Psi_{22}) = (16, 20)$ corresponds to a moderate correlation (between 0.80 and 0.90).

To estimate the coverage probability of a UCB at a given setting, we simulate realizations from the distribution (12), fit the model using maximum likelihood, and compute the UCB as described in Section 3 for the no-covariate case. This process is repeated 2500 times when $t_{m-l}(\alpha)$ is used as the critical point and 1000 times with $B = 500$ when the bootstrap-$t$ critical point is used. The proportion of times a UCB is correct gives its estimated coverage probability. They are reported in Table 1 for the UCB $U_x$ of the percentile $q_x$, given by (8); and in Table 2 for the UCB $U_{x2}$ of the percentile $q_{x2}$, given by (10). The results for $q_{x1}$ are not presented separately as $q_{x1} = q_{x2}$ when $(\beta_1, \lambda_1, \Psi_{11}) = (\beta_2, \lambda_2, \Psi_{22})$. The coverage probability of $U_{x2}$ is free of $p_0$ since the term involving it in $q_{x2}$ is a known constant.

From Table 1, we conclude that $U_x$ with $t_{m-l}(\alpha)$ as the critical point is slightly conservative when $\Psi_{11} = \Psi_{22}$—its coverage probability estimates are about 1% higher than the target nominal level of 95%. On the other hand, it is liberal

Table 1
Estimated coverage probabilities (%) of the 95% confidence level UCB $U_x$ computed using (8) for the no-covariate case

| $(\beta_2, \lambda_2, \Psi_{22})$ | $c_{0.05} = t_{m-2}(0.05)$ | | | | | | Bootstrap $c_{0.05}$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $m = 15$ | | | $m = 30$ | | | $m = 15$ | | | $m = 30$ | | |
| | $n$ | | | | | | $n$ | | | | | |
| | 2 | 3 | 5 | 2 | 3 | 5 | 2 | 3 | 5 | 2 | 3 | 5 |
| (0, 1.0, 16) | 96.1 | 97.2 | 96.8 | 96.3 | 96.2 | 95.2 | 94.4 | 93.4 | 93.8 | 95.3 | 93.9 | 93.7 |
| (2, 1.0, 16) | 95.6 | 95.7 | 96.1 | 94.9 | 95.3 | 94.8 | 95.5 | 96.2 | 94.8 | 96.2 | 95.3 | 94.2 |
| (0, 1.5, 16) | 96.2 | 96.2 | 96.8 | 95.9 | 96.0 | 95.5 | 95.9 | 93.8 | 95.2 | 94.6 | 96.0 | 94.1 |
| (2, 1.5, 16) | 96.1 | 96.0 | 96.6 | 95.8 | 96.0 | 94.8 | 95.9 | 95.4 | 94.1 | 95.2 | 94.0 | 95.6 |
| (0, 1.0, 20) | 91.8 | 91.6 | 91.9 | 92.8 | 91.6 | 92.1 | 93.8 | 93.9 | 94.4 | 94.0 | 94.8 | 94.7 |
| (2, 1.0, 20) | 92.0 | 92.4 | 92.4 | 93.2 | 93.4 | 92.9 | 93.9 | 94.7 | 95.5 | 94.5 | 95.6 | 94.6 |
| (0, 1.5, 20) | 93.0 | 91.9 | 91.1 | 92.7 | 92.4 | 91.7 | 93.8 | 94.1 | 93.8 | 94.9 | 94.7 | 92.3 |
| (2, 1.5, 20) | 92.5 | 93.0 | 92.2 | 93.6 | 92.9 | 93.1 | 94.0 | 93.4 | 94.6 | 94.1 | 95.3 | 96.3 |

Throughout we have $(p_0, \beta_1, \lambda_1, \Psi_{11}, \Psi_{12}) = (0.80, 0, 1, 16, 15.95)$. The estimates in the left half are based on 2500 replications and have a standard error of 0.4%. The estimates in the right half are based on 1000 replications and have a standard error of 0.7%.

Table 2
Estimated coverage probabilities (%) of the 95% confidence level UCB $U_{x2}$ computed using (10) for the no-covariate case

| $(\beta_2, \lambda_2, \Psi_{22})$ | $c_{0.05} = t_{m-2}(0.05)$ | | | | | | Bootstrap $c_{0.05}$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $m = 15$ | | | $m = 30$ | | | $m = 15$ | | | $m = 30$ | | |
| | $n$ | | | | | | $n$ | | | | | |
| | 2 | 3 | 5 | 2 | 3 | 5 | 2 | 3 | 5 | 2 | 3 | 5 |
| (0, 1.0, 16) | 90.8 | 93.1 | 93.9 | 92.1 | 92.9 | 94.0 | 93.5 | 94.3 | 93.1 | 94.5 | 93.8 | 95.2 |
| (2, 1.0, 16) | 90.5 | 92.0 | 93.0 | 91.6 | 92.0 | 93.1 | 93.9 | 93.7 | 93.9 | 94.3 | 94.5 | 93.6 |
| (0, 1.5, 16) | 89.8 | 91.1 | 93.7 | 90.3 | 92.7 | 93.7 | 95.4 | 93.8 | 94.8 | 93.4 | 95.5 | 94.5 |
| (2, 1.5, 16) | 90.0 | 91.9 | 93.1 | 90.8 | 92.0 | 93.5 | 93.9 | 95.9 | 95.5 | 93.6 | 93.8 | 95.0 |
| (0, 1.0, 20) | 92.2 | 94.5 | 93.7 | 93.2 | 92.5 | 93.8 | 95.6 | 94.8 | 93.5 | 94.3 | 94.1 | 94.3 |
| (2, 1.0, 20) | 92.3 | 94.6 | 94.6 | 92.8 | 93.1 | 93.6 | 94.6 | 95.3 | 94.8 | 94.4 | 95.2 | 94.0 |
| (0, 1.5, 20) | 92.4 | 93.5 | 94.4 | 93.5 | 93.2 | 94.2 | 95.1 | 93.9 | 95.4 | 95.3 | 95.2 | 95.2 |
| (2, 1.5, 20) | 92.0 | 93.4 | 95.0 | 93.0 | 94.2 | 94.2 | 95.3 | 94.6 | 95.2 | 93.6 | 94.2 | 94.7 |

These probabilities do not depend on $p_0$. Throughout we have $(\beta_1, \lambda_1, \Psi_{11}, \Psi_{12}) = (0, 1, 16, 15.95)$. The estimates in the left half are based on 2500 replications and have a standard error of 0.4%. The estimates in the right half are based on 1000 replications and have a standard error of 0.7%.

when $\Psi_{11} \neq \Psi_{22}$—the estimates are lower than the target. They are about 3% lower in case of $m = 15$ and increase by about 0.5–1% for $m = 30$. The estimates remain more or less similar across different values of $\beta_2$ and $\lambda_2$. Surprisingly $n$ also does not seem to have an impact. Additional investigation reveals that the normality assumption for $\log \hat{q}_x$ is reasonable and there is no evidence of any substantial bias in the estimation. However, when $\Psi_{11} = \Psi_{22}$, $var(\log \hat{q}_x)$ tends to get overestimated, while the converse is true when $\Psi_{11} \neq \Psi_{22}$. This seems to be the cause of the conservative behavior of $U_x$ in the former case and its liberal behavior in the latter case. Fortunately, this problem can be resolved by using the bootstrap critical point in (8) for $U_x$. The resulting UCB is quite accurate even with $m = 15$ irrespective of whether or not $\Psi_{11} = \Psi_{22}$ as all the probability estimates in the right half of Table 1 are near 0.95.

Table 2 suggests that $U_{x2}$ with $t_{m-l}(\alpha)$ is liberal throughout. In contrast with $U_x$ above, this liberal behavior is more severe when $\Psi_{11} = \Psi_{22}$ than when $\Psi_{11} \neq \Psi_{22}$; and $U_{x2}$ becomes more accurate as $n$ increases. The increase in accuracy with an increased $m$ is expected. These estimates also remain somewhat constant as $\beta_2$ and $\lambda_2$ vary. In this case also, using bootstrap leads to fairly accurate bounds at all settings.

Additional simulations with $m = 60$ indicate that the coverage probabilities of the bounds with $t_{m-l}(\alpha)$ increase by 1–1.5% over $m = 30$ to 93–94% in regions where they are liberal, and remain close to the target 95% in regions

where they are conservative. Overall, due to its remarkable accuracy, the bootstrap-$t$ approach is recommended for constructing UCBs, particularly when $m \leqslant 60$. A downside of this approach is that it is computationally demanding. It took about 16 min to compute a bootstrap-$t$ critical point for $m = 15$ and about 28 min for $m = 30$, on a Dell laptop with a 1.8 GHz Pentium 4 processor, 512 MB of RAM and Windows XP operating system. However, once we have a function for fitting mixed models, bootstrapping becomes a routine computation. And almost all the popular statistical packages in current usage, including R, have this capability.

**Remark**: Sometimes it may be of interest to impose some structure on the distribution of $(b_{i1}, b_{i2})$. Two choices seem popular in the literature. The first one assumes $b_{ij} = b_i + b_{i*j}$, where $b_i$ is the true unobservable measurement for the $i$th individual; $b_{i*j}$ is the method-individual interaction; and $(b_i, b_{i*1}, b_{i*2})$ are mutually independent normal random variables with different variances (see e.g., Bland and Altman, 1999). This structure can be easily incorporated in (11) by taking $\Psi_{jj} = var(b_i) + var(b_{i*j})$ and $\Psi_{12} = var(b_i)$. In practice, however, there is generally no difference between letting $b_{ij}$'s in (11) be arbitrary or imposing this structure since $\Psi_{12} = cov(b_{i1}, b_{i2})$ in (12) tends to be positive. The second choice is to assume a linear relationship, say, $\beta_2 + b_{i2} = \alpha_1 + \alpha_2(\beta_1 + b_{i1})$, between $b_{i1}$ and $b_{i2}$ (see e.g., Dunn and Roberts, 1999). The resulting model is a *structural equation model* and it cannot be written as a mixed model.

## 5. Application

### 5.1. Cardiac output data

The exploratory analysis by Bland and Altman (1999) suggests the no-covariate model (11) for these data. This model seems to fit well as indicated by the residual plot in Fig. 1 and other diagnostic plots recommended by Pinheiro and Bates (2000, Chapter 4). Identical fit results if we replace $b_{ij}$ in (11) with $b_i + b_{i*j}$, where $b_i$ represents the random individual effect and $b_{i*j}$ is the random method-individual interaction. We have the following MLEs of the parameters in (12), with their standard errors in parentheses: $\hat{\beta}_1 = 5.39(0.37)$, $\hat{\beta}_2 = 4.68(0.35)$, $\hat{\Psi}_{11} = 1.63(0.68)$, $\hat{\Psi}_{12} = 1.15(0.56)$, $\hat{\Psi}_{22} = 1.45(0.60)$, $\hat{\lambda}_1 = 0.11(0.02)$ and $\hat{\lambda}_2 = 0.14(0.03)$. Thus, the estimated mean and standard deviation of the population of RV measurements are 5.39 and 1.32, respectively; and they are 4.68 and 1.26 for the population of IC measurements. Also, their estimated correlation is 0.69.

The estimates of intra-class correlation for RV and IC methods are 0.94 and 0.91, respectively. Further, the estimated standard deviation of the population difference between any two RV measurements is 0.47 and is 0.53 for the IC measurements. Taking $(p_0, \alpha) = (0.80, 0.05)$ and using (10) for the no-covariate case, we get the tolerance intervals for the intra-method agreement as $[-0.71, 0.71]$ (RV) and $[-0.81, 0.81]$ (IC). The bootstrap-$t$ approach produces similar intervals: $[-0.70, 0.70]$ and $[-0.81, 0.81]$. Thus, 80% of IC differences are estimated to lie within $\pm 0.81$, and a similar interpretation holds for RV differences. These findings confirm that both methods have good repeatability and RV is slightly more repeatable than IC.

Next, we consider the agreement between the methods. From (6), the population of RV–IC differences has estimated mean and standard deviation of 0.70 and 1.01, respectively. Upon taking $(p_0, \alpha) = (0.80, 0.05)$ and $t_{10}(0.05)$ as the critical point in (8), we get $[-2.18, 2.18]$ as the tolerance interval. This interval widens to $[-2.33, 2.33]$ when we use the bootstrap-$t$ critical point. The latter interval is probably more accurate since $m = 12$ is not large and the correlation 0.69 is weak. Thus, 80% of RV–IC differences are estimated to lie in $[-2.33, 2.33]$. To infer whether this extent of agreement is sufficient, one can compare it with a threshold interval provided by the practitioner such that the differences in the interval are deemed clinically unimportant. When such a threshold is not explicitly available, a practical strategy is to compare the bound $U_x$ with the magnitude of measurements—if it is large relative to the magnitude, one infers insufficient agreement; otherwise sufficient agreement is inferred. This strategy is effective when the bound is either quite large or quite small relative to the magnitude making the conclusion straightforward. In real applications, one may take the average measurement as a proxy for the unknown true magnitude of measurement. Another strategy may be to compare the bound with the range of measurement. For the cardiac output data, the measurements range between 2 and 8, and $U_x = 2.33$ is approximately 45% of the average measurement of about 5.0—indicating poor agreement in RV and IC methods.

### 5.2. Body fat data

In this example, we directly model the time profiles of the differences (caliper $y_1$ − DEXA$y_2$) of the paired body fat measurements. A preliminary analysis suggests modelling the mean time profiles nonparametrically as a quadratic spline in $x =$ (age in years − 12) $\in \mathfrak{X} = [-0.80, 5.30]$,

$$f(x, \boldsymbol{\beta}, \mathbf{u}) = \beta_0 + \beta_1 x + \beta_2 x^2 + \sum_{j=1}^{r} u_j (x - \kappa_j)_+^2, \tag{13}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$, $\mathbf{u} = (u_1, \ldots, u_r)$, $\kappa_1 < \cdots < \kappa_r$ are the knot locations in $\mathfrak{X}$; $u_1, \ldots, u_r$ are the coefficients of the truncated quadratic basis functions $(x - \kappa_1)_+^2, \ldots, (x - \kappa_r)_+^2$, and $(x - \kappa)_+ = \max\{0, x - \kappa\}$. See Ruppert et al. (2003, Chapters 3–6) for an excellent introduction to penalized splines regression. The use of *quadratic* basis functions in agreement applications is suggested by Choudhary and Ng (2006). The number of knots $r = 34$ and their locations $\kappa_j = ((j+1)/(r+2))$th sample percentile of the unique observed values of $x \in \mathfrak{X}$, $j = 1, \ldots, r$, are chosen using the recommendation of Ruppert et al. (2003, Chapter 5).

The model that we fit to the body fat differences is,

$$d_{ik} = f(x_{ik}, \boldsymbol{\beta}, \mathbf{u}) + b_i + \varepsilon_{ik}; \quad k = 1, \ldots, n_i, \ i = 1, \ldots, m,$$

where $x_{ik}$ is the value of $x$ at the $k$th visit of the $i$th individual; $d_{ik}$ is the difference associated with $x_{ik}$; $f$ is given in (13); $b_i$ is the random-effect of the $i$th individual; and $\varepsilon_{ik}$ is the within-individual random error. We assume that $b_i \sim \mathcal{N}(0, \Psi)$, independently of errors that follow mean zero normal distributions with $cov(\varepsilon_{ik}, \varepsilon_{il}) = \lambda_1 \lambda_2^{|x_{ik} - x_{il}|}$ $(k, l = 1, \ldots, n_i)$. Here, $\lambda_1$ is the within-individual error variance and $\lambda_2$ is the non-negative correlation between two within-individual errors one unit of time apart. This covariance structure is equivalent to assuming that the within-individual errors follow a continuous autoregressive process of order one (see, e.g., Pinheiro and Bates, 2000, Chapter 5). To fit the spline (13), we use the penalized criterion, which is equivalent to assuming that $u_1, \ldots, u_r$ follow independent $\mathcal{N}(0, \Psi_u)$ distributions, mutually independent of the random intercepts and errors (see, e.g., Ruppert et al., 2003, Chapter 4). Thus in this case, the vector $\mathbf{d}_i$ of $n_i$ differences on the $i$th individual, $i = 1, \ldots, m$, is modelled as

$$\mathbf{d}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i b_i + \mathbf{W}_i \mathbf{u} + \boldsymbol{\epsilon}_i; \quad b_i \sim \mathcal{N}(0, \Psi), \ \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \Psi_u \mathbf{J}_r), \ \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_i) \tag{14}$$

with $\mathbf{X}_i$ as the $n_i \times 3$ matrix with $k$th row $(1, x_{ik}, x_{ik}^2)$; $\mathbf{Z}_i$ as the $n_i$-vector of 1's; $\mathbf{W}_i$ as the $n_i \times r$ matrix with $k$th row $((x_{ik} - \kappa_1)_+^2, \ldots, (x_{ik} - \kappa_r)_+^2)$; $\mathbf{J}_r$ as the $r \times r$ identity matrix; and $\boldsymbol{\Lambda}_i$ as the $n_i \times n_i$ matrix with $(k, l)$th element $\lambda_1 \lambda_2^{|x_{ik} - x_{il}|}$. This mixed model has two levels of independent random-effects—the random intercept $b_i$ that varies with individuals and the knot coefficient vector $\mathbf{u}$ that is common to all individuals. They are also mutually independent of error $\boldsymbol{\epsilon}_i$. All these quantities are independent for different $i$. When $\Psi_u = 0$, the $\mathbf{u}$ term in (14) vanishes and the model reduces to (5). We will fit this model using maximum likelihood.

As in Section 3, the evaluation of agreement here will focus on the distribution of population difference $d_x$ at $x$. Averaging over the random individual effects in (14) leads to

$$d_x \sim \mathcal{N}(\mu_x = f(x, \boldsymbol{\beta}, \mathbf{u}), \sigma^2 = \Psi + \lambda_1), \quad x \in \mathfrak{X}.$$

Substitution in (7) gives $q_x$—the measure of agreement at $x$. It is now a random parameter since it involves the random $\mathbf{u}$. To estimate it, we replace $(\boldsymbol{\beta}, \sigma)$ in $q_x$ with its MLE and $\mathbf{u}$ with its estimated best linear unbiased predictor (EBLUP) $\hat{\mathbf{u}}$ computed as described in Pinheiro and Bates (2000, Chapter 2). However, due to the random nature of $q_x$, we cannot directly use (8) to obtain its simultaneous UCB. We now generalize the methodology of Choudhary and Ng (2006) to get an approximation that is expected to work well when $\Psi_u/\sigma^2$ is small. (Its MLE for the body fat data is 0.05.) Let $\xi$ denote the $(r+4) \times 1$ column vector $(\boldsymbol{\beta}, \mathbf{u}, \log \sigma)$, $\hat{\xi} = (\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}, \log \hat{\sigma})$ be its estimate, $V$ be the asymptotic covariance matrix of $(\hat{\xi} - \xi)$ and $\mathbf{G}_x = (\partial \log q_x / \partial \xi)_{\xi = \hat{\xi}}$. When $m$ is large and $\Psi_u/\sigma^2$ is small, we expect $\log \hat{q}_x - \log q_x \approx \mathcal{N}(0, \mathbf{G}_x' \mathbf{V} \mathbf{G}_x)$. So this UCB of $q_x$ is also of the form, $U_x = \exp\{\log \hat{q}_x - c_\alpha (\mathbf{G}_x' \mathbf{V} \mathbf{G}_x)^{1/2}\}$, where the critical point $c_\alpha$ is computed by solving (9), but $V$ is now estimated using bootstrap in the following manner:

1. Take a fine grid $x_1, \ldots x_g$ of equally spaced points in $\mathfrak{X}$. Say $g = 100$.

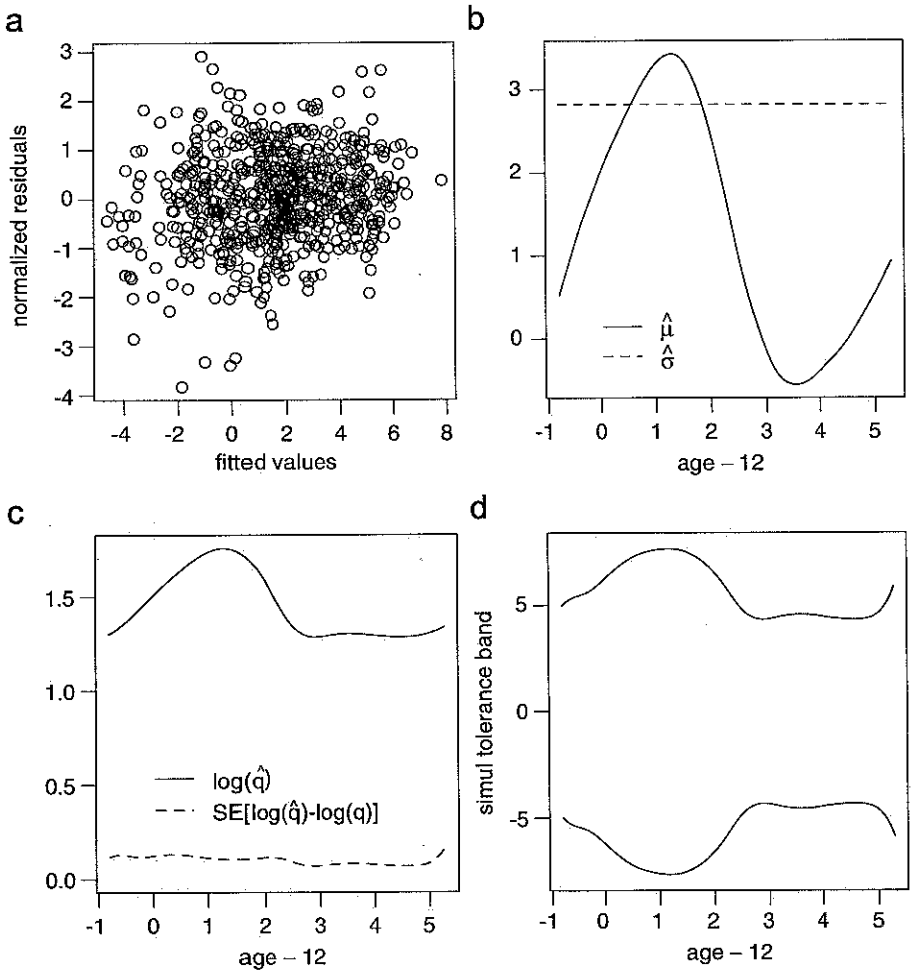


Fig. 3. (a) Residual plot of the model fitted to the body fat data. (b) Estimated mean function and standard deviation of the population difference $d_x$. (c) Estimated $\log q_x$ function with $p_0 = 0.80$ and the standard error of $(\log \hat{q}_x - \log q_x)$. (d) The 80% tolerance band for the distribution of $d_x$ with simultaneous 95% confidence.

2. Generate independent $u_1^*, \ldots, u_r^*$ from $\mathcal{N}(0, \hat{\Psi}_u)$ and set $\xi^* = (\hat{\beta}, \mathbf{u}^*, \log \hat{\sigma})$. Here $\hat{\Psi}_u$ is the MLE of $\Psi_u$ and $\mathbf{u}^* = (u_1^*, \ldots, u_r^*)$.
3. Generate $d_{x_i}^*$ independently from $\mathcal{N}(0, f(x_i, \hat{\beta}, \hat{\mathbf{u}}), \hat{\sigma}^2)$, $i = 1, \ldots, g$. These $(x_i, d_{x_i}^*)$ pairs represent a parametric bootstrap resample.
4. Fit the model $d_{x_i}^* \sim$ independent $\mathcal{N}(0, f(x_i, \beta, \mathbf{u}^*), \sigma^2)$ to the above resample using maximum likelihood and obtain the MLE of $(\beta, \sigma)$, say $(\hat{\beta}^*, \hat{\sigma}^*)$; the EBLUP of $\mathbf{u}^*$, say $\hat{\mathbf{u}}^*$; set $\xi^* = (\hat{\beta}^*, \hat{\mathbf{u}}^*, \log \hat{\sigma}^*)$; and compute $(\xi^* - \xi^*)$.
5. Repeat steps 2–4 a large number of times, say $B = 500$, to obtain $B$ realizations $(\xi^* - \xi^*)$, and use their sample covariance matrix as an estimate of $V$.

We get the following MLEs upon fitting the model (14) to the body fat data: $\hat{\beta}_0 = 2.03(0.54)$, $\hat{\beta}_1 = 1.59(1.18)$, $\hat{\beta}_2 = -0.36(1.15)$, $\hat{\Psi} = 4.38(0.83)$, $\hat{\Psi}_u = 0.40(0.30)$, $\hat{\lambda}_1 = 3.59(0.35)$ and $\hat{\lambda}_2 = 0.21(0.05)$. The parentheses contain the standard errors. The fitted time profiles are presented in Fig. 2 and the plot of normalized residuals is given in Fig. 3(a). These indicate a reasonable model fit, although there are a few high residuals. The normality assumption for the random-effects and the errors also seems fine. Further, a likelihood ratio test of significance for $\Psi_u$ has a $p$-value $< 10^{-4}$, which justifies the $\mathbf{u}$ term in (14). Thus overall, the fitted model appears appropriate. The resulting estimated mean and standard deviation of $d_x$ over $\mathfrak{X}$ are given in panel (b) of Fig. 3. The graph of the fitted mean function is consistent with the shape of the observed time profiles. The panel (c) of this figure presents the estimated $\log q_x$ function

for $p_0 = 0.80$ and the standard error of $(\log \hat{q}_x - \log q_x)$. The last panel plots the 80% tolerance band $[-U_x, U_x]$ for the distribution of $d_x$ with simultaneous 95% confidence, i.e., (1) holds with $(p_0, \alpha) = (0.80, 0.05)$. This band demonstrates the estimated range of differences in 80% of body fat measurements from the two methods as a function of age of the girls. The agreement between the methods appears best around age 15–17 where 80% of the measurements can differ as much as by about 4.5%. In this region, the magnitude of measurements, as suggested by the average measurement, is about 25%. On the whole, the agreement between the skinfold calipers and DEXA methods does not seem good enough to justify their interchangeable use since a change of 3–6% in percentage body fat measurements is considered important as it may lead from one category of body fat, such as essential fat, athletes, fitness, acceptable and obesity, to another.

## 6. Discussion

In this paper, we extended the tolerance interval approach for assessing agreement between two methods of measurement to deal with the repeated measurement data, assuming that the observed data can be modelled using a linear mixed model. Besides the tolerance intervals, there are several other measures of agreement—including the limits of agreement of Bland and Altman (1986) and the concordance correlation of Lin (1989). See the reviews of Lin et al. (2002) and Choudhary and Nagaraja (2004) for a comparison of measures. Here we just note that the basic limits of agreement have been generalized by Bland and Altman (1999) for the model (11). Furthermore, using a random-coefficient growth curve model (a special case of a mixed model), Chinchilli et al. (1996) have proposed a weighted average of individual-specific concordance correlations as a single overall measure of agreement for repeated measurement data. However, their individual-specific concordance correlations quantify the agreement between a linear transformation (e.g., sample means) of the observed responses from the two methods, whereas we have been concerned with measuring agreement between the individual responses. Recently, Barnhart et al. (2005) provide another extension of the concordance correlation for repeated measurement data. In addition to the measure of agreement being used, our approach differs from these extensions in two respects: first, we use a general mixed model framework; and second, we allow the extent of agreement to depend on a continuous covariate. It also appears possible to adapt the approach of this paper to extend these other agreement measures.

An attractive alternative to mixed models for modelling dependent data is the framework of *marginal models* (see, e.g., Diggle et al., 2002, Chapter 7). In a marginal model, one specifies separate models for the marginal mean of response, the marginal variance of response and the within-individual correlation in response. There is no need to specify the entire likelihood as one uses the *generalized estimating equations* approach for inference. In contrast, in a mixed model, the within-individual correlation is induced by common random effects in the model for response, and it generally requires full specification of the likelihood function. Marginal models are appropriate when inference about the mean response, or more generally, a function of the moments is of primary interest. Indeed, Barnhart and Williamson (2001) and Barnhart et al. (2005) have used them successfully for agreement assessment with concordance correlation, which is a function of the first two moments. But in the tolerance interval approach, a percentile is the main focus of inference, and without additional assumptions regarding the distribution of response, the moments do not determine a percentile. However, as this paper demonstrated, this inference is straightforward in a mixed model setup.

Finally, we note that we have not addressed the important issue of how to design a method comparison study. This involves determining the number of individuals and the number of replicate measurements in some optimal fashion. Further research is needed in this direction.

## Acknowledgments

The author thanks Prof. V.M. Chinchilli, Prof. T.S. King and Prof. J.M. Bland for providing the data sets, and Prof. Tony Ng for helpful discussions. He is also thankful to the reviewers whose comments led to substantial improvements in this paper and to the Executive Editor Prof. John Stufken for his consideration.

## References

Barnhart, H.X., Williamson, J.M., 2001. Modeling concordance correlation via GEE to evaluate reproducibility. Biometrics 57, 931–940.
Barnhart, H.X., Song, J., Haber, M.J., 2005. Assessing intra, inter and total agreement with replicated readings. Statist. Med. 24, 1371–1384.

Bland, J.M., Altman, D.G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet i, 307–310.
Bland, J.M., Altman, D.G., 1999. Measuring agreement in method comparison studies. Statist. Methods Med. Res. 8, 135–160.
Chinchilli, V.M., Martel, J.K., Kumanyika, S., Lloyd, T., 1996. A weighted concordance correlation coefficient for repeated measurement designs. Biometrics 52, 341–353.
Choudhary, P.K., Nagaraja, H.N., 2004. Measuring agreement in method comparison studies—a review. In: Balakrishnan, N., Kannan, N., Nagaraja, H.N. (Eds.), Advances in Ranking and Selection, Multiple Comparisons, and Reliability. Birkhauser, Boston, pp. 215–244.
Choudhary, P.K., Nagaraja, H.N., 2007. Tests for assessment of agreement using probability criteria. J. Statist. Plann. Inference 137, 279–290.
Choudhary, P.K., Ng, H.K.T., 2006. Assessment of agreement under non-standard conditions using regression models for mean and variance. Biometrics 62, 288–296.
Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and Their Application. Cambridge University Press, New York.
Diggle, P.J., Heagertry, P., Liang, K.-L., Zeger, S.L., 2002. Analysis of Longitudinal Data. second ed. Oxford University Press, New York.
Dunn, G., Roberts, C., 1999. Modelling method comparison data. Statist. Methods Med. Res. 8, 161–179.
Fleiss, J.L., 1986. The Design and Analysis of Clinical Experiments. Wiley, New York.
Hawkins, D.M., 2002. Diagnostics for conformity of paired quantitative measurements. Statist. Med. 21, 1913–1935.
Lin, L.I., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255–268 (Corrections: 2000, 56:324-325).
Lin, L.I., 2000. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. Statist. Med. 19, 255–270.
Lin, L.I., Hedayat, A.S., Sinha, B., Yang, M., 2002. Statistical methods in assessing agreement: models, issues, and tools. J. Amer. Statist. Assoc. 97, 257–270.
Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-PLUS. Springer, New York.
R Development Core Team, 2004. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. {http://www.R-project.org}.
Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression. Cambridge University Press, New York.