Measuring Agreement in Method Comparison Studies With Heteroscedastic Measurements

Lakshika S. Nawarathna and Pankaj K. Choudhary[1]

Department of Mathematical Sciences, FO 35

University of Texas at Dallas

Richardson, TX 75083-0688, USA

Abstract

We propose a methodology for evaluation of agreement between two methods of measuring a continuous variable whose variability changes with magnitude. This problem routinely arises in method comparison studies that are common in health-related disciplines. Assuming replicated measurements, our approach is to first model the data using a heteroscedastic mixed-effects model, wherein a suitably defined true measurement serves as the variance covariate. Fitting this model poses some computational difficulties as the likelihood function is not available in a closed form. We deal with this issue by suggesting four estimation methods to get approximate maximum likelihood estimates. Two of these methods are based on numerical approximation of the likelihood and the other two are based on approximation of the model. Next, we extend the existing agreement evaluation methodology designed for homoscedastic data to work under the proposed heteroscedastic model. This methodology can be used with any scalar measure of agreement. Simulations show that the suggested inference procedures generally work well for moderately large samples. They are illustrated by analyzing a data set of cholesterol measurements.

Keywords: Concordance correlation, generalized linear mixed-effects models, limits of agreement, nonlinear mixed-effects model, total deviation index.

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

1 Introduction

Hundreds of method comparison studies are published each year in health-related fields such as medicine, biomedical engineering, medical imaging, nutrition and clinical chemistry. These studies compare a new cheaper, simpler or less invasive method of measuring a continuous variable with an established method to see if they have sufficient agreement for interchangeable use. The methods may be assays, clinical observers, medical devices, etc. The variable of interest typically has some clinical importance, e.g., concentration of a chemical, blood pressure, cholesterol level, etc. Each subject in the study is measured at least once by every method. The statistical methodology for evaluation of agreement in such studies is well developed for the case when the variability of measurement remains constant over the entire measurement range. Reviews of the literature on this topic can be found in [1–4].

This methodology generally consists of two steps. The first step is to model the method comparison data. It is common to use a *mixed-effects model* [5] assuming normality for both random effects and errors, and with variances that may depend on the method but remain constant over the measurement range [6–11]. The second step is to evaluate agreement between the methods by performing inference on one or more *measures of agreement* that quantify how well the methods agree. Essentially good agreement between two methods means small differences in their measurements. The agreement measures are appropriate functions of parameters of the model fitted in the first step. A number of such measures are available in the literature, see [1] for a detailed review.

In practice, the error variances of the methods often depend on the magnitude of measurement [12–14], violating the homoscedasticity assumption of the model. A real example of this is the cholesterol data from [15]. These data come from a trial involving 100 subjects in which serum cholesterol (mg/dl) is measured ten times using each of two assays — Cobas Bio (method 1) and Ekatachem 700 (method 2). The first assay is a Centers for Disease

Control standardized method serving as a reference, whereas the second assay is a routine laboratory analyzer serving as a test method. The design of these data is balanced and there is a total of $100 \times 2 \times 10 = 2000$ observations. We are interested in quantifying the extent of agreement between the two assays to see if they can be used interchangeably. Figure 1 shows a trellis plot of the data. Although there is considerable overlap between the two assays, the Ektachem measurements tend to be larger and have higher within-subject variation than the Cobas Bio measurements. Moreover, for both assays, this variation seems to increase with the magnitude of measurement but it remains substantially lower than the between-subject variation. There is also evidence of assay × subject interaction. Initially we fit the usual homoscedastic mixed-effects model — to be presented in Section 2.1 — to these data via maximum likelihood (ML) using the `nlme` package [16] in `R` [17]. Figure 2 shows the resulting residual plot for each assay. The fan-shaped pattern in both plots confirms that the error variation of each assay increases with the magnitude of measurement.

Unlike the homoscedastic case, the extent of agreement in the heteroscedastic case is not constant because it depends on the magnitude of measurement. Thus, if this heteroscedasticity is not taken into account at the data modeling stage, the subsequent agreement evaluation could be misleading as it would mistakenly treat the agreement measure to be a constant. It may be possible to remove the heteroscedasticity by a variance stabilizing transformation of data [13, 14], but a transformation is not always successful. Besides, a transformation other than log is generally not recommended in method comparison studies because the differences of the transformed measurements may be difficult to interpret [13].

The goal of this article is to develop a methodology for agreement evaluation when the error variances of the two methods change with magnitude of measurement. Instead of removing the heteroscedasticity altogether, we explicitly model it in the data modeling step using a suitably defined true measurement as the *variance covariate* [5, ch. 5]. Such a model is a heteroscedastic mixed-effects model [5, 18]. We then generalize the agreement evaluation methodology designed for homoscedastic data to work under this heteroscedastic model. This methodology can be used with any scalar measure of agreement currently available in the literature. We assume that the measurements are replicated, i.e., each subject in the study is measured more than once by every method under identical conditions. However, the design of the study need not be balanced.

Note that the true measurement in method comparison studies, and hence our variance covariate, is an unobservable random quantity. This contrasts with variance covariates in heteroscedastic regression models [19] wherein they are possibly unknown but non-random quantities. Further, our approach differs from [20] which models differences in *unreplicated* measurements from two methods. Here we model all data, not just the differences.

This article is organized as follows. In Section 2, we describe a heteroscedastic mixed-effects model for method comparison data and discuss its fitting. We also provide tests for heteroscedasticity in this section. In Section 3, we use simulation to examine properties of the proposed estimation and testing procedures. In Section 4, we extend the existing agreement evaluation methodology developed for homoscedastic models to work with the heteroscedastic model. In Section 5, we illustrate the methodology by analyzing the cholesterol data introduced earlier in this section. We conclude in Section 6 with a discussion. The statistical software `R` [17] has been used for all the computations in this article.

2 A heteroscedastic model for method comparison data

Consider a method comparison data set consisting of Y_{ijk}, $k = 1, \ldots, n_{ij}$ (≥ 2), $j = 1, 2$, $i = 1, \ldots, m$, where Y_{ijk} is the kth replicate measurement by the jth method on the ith subject. Here m is the number of subjects in the study and n_{ij} is the number of measurements from the jth method on the ith subject. Let $n_i = n_{i1} + n_{i2}$ be the total number of measurements on the ith subject. It is assumed that the multiple measurements on a subject made by the same method are replications of the same true underlying measurement. Moreover, the replicates from the two methods are not paired in that their time ordering is immaterial.

Let $\mathbf{Y}_{ij}$ be a $n_{ij} \times 1$ vector of measurements $(Y_{ij1}, \ldots, Y_{ijn_{ij}})$ on the ith subject from the jth method. Also let $\overline{Y}_{ij} = \sum_{k=1}^{n_{ij}} Y_{ijk}/n_{ij}$ be the sample mean of these measurements. The $n_i \times 1$ vector $\mathbf{Y}_i = (\mathbf{Y}_{i1}, \mathbf{Y}_{i2})$ denotes all the data on ith subject, and the 2×1 vector $\overline{\mathbf{Y}}_i = (\overline{Y}_{i1}, \overline{Y}_{i2})$ denotes the vector of the sample means. All vectors in this article are column vectors unless specified otherwise. The vectors and matrices are bold-faced. The transpose of a vector or matrix $\mathbf{A}$ is denoted as $\mathbf{A}^T$. We will use $\boldsymbol{\theta}$ to denote the vector of all unknown model parameters. We will also use $f_{\boldsymbol{\theta}}(\mathbf{y}_1, \mathbf{y}_2)$ for joint probability density function of $(\mathbf{Y}_1, \mathbf{Y}_2)$, and $f_{\boldsymbol{\theta}}(\mathbf{y}_1|\mathbf{y}_2)$ for conditional probability density function of $\mathbf{Y}_1|\mathbf{Y}_2 = \mathbf{y}_2$.

2.1 A common homoscedastic mixed-effects model

It is common to assume that the method comparison data follow the mixed-effects model:

$$Y_{ijk} = \beta_j + b_{ij} + e_{ijk}, \quad k = 1, \ldots, n_{ij}, \quad j = 1, 2, \quad i = 1, \ldots, m, \tag{1}$$

where β_j is the fixed mean of the jth method, b_{ij} is the random effect of the ith subject on the jth method and e_{ijk} is the within-subject random error. Let $\mathbf{b}_i = (b_{i1}, b_{i2})$ be the 2×1 vector of random effects of the ith subject. It is assumed that the errors and the random effects follow mutually independent normal distributions,

$$e_{ijk} \sim \text{ independent } \mathcal{N}_1\big(0, \sigma_j^2\big), \quad \mathbf{b}_i \sim \text{ independent } \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Psi}), \tag{2}$$

where $\boldsymbol{\Psi}$ is a 2×2 positive definite matrix with ψ_1^2 and ψ_2^2 as diagonal elements, ψ_{12} as the off-diagonal element, and $\rho = \psi_{12}/(\psi_1\psi_2)$ as the correlation. This homoscedastic model has been used in several articles, including [8, 11, 21].

Let $\mu_{ij} = \beta_j + b_{ij}$ be the conditional mean, $E(Y_{ijk}|\mathbf{b}_i)$. It represents the unobservable "true" (i.e., error-free) measurement of the jth method on the ith subject under model (1).

The true measurements of the methods may be unequal due to the method × subject interaction and a difference in the means. Next, let $\boldsymbol{\beta} = (\beta_1, \beta_2)$ be the 2×1 vector of the fixed means and $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}) = \boldsymbol{\beta} + \mathbf{b}_i$ be the 2×1 vector of the true values.

2.2 The proposed heteroscedastic mixed-effects model

Let v be a variance covariate and v_i be its value for the ith subject. To model the error variation as a function of the magnitude of measurement, we would like v to be the true measurement. But the absolute truth is not available in method comparison studies. Therefore, as in [14], one practical alternative is to take $v_i = \mu_{i1}$ if one of the methods in the comparison, say, method 1, is an established standard method serving as a reference, otherwise, $v_i = (\mu_{i1} + \mu_{i2})/2$. In either case, v_i is a function of $\boldsymbol{\mu}_i$, say, $v_i = h(\boldsymbol{\mu}_i)$, and is an unobservable random quantity serving as a proxy for the true measurement.

Next, let $g(v, \boldsymbol{\delta})$ denote a *variance function* — a function of v describing how the variation changes with v. This function has a known form but may involve an unknown heteroscedasticity parameter vector $\boldsymbol{\delta}$ such that $g(v, \boldsymbol{\delta}) \equiv 1$ when $\boldsymbol{\delta} = \mathbf{0}$, corresponding to homoscedasticity. Some common variance function models include $g(v, \delta) = |v|^{\delta}$ (power model), $g(v, \boldsymbol{\delta}) = \delta_0 + |v|^{\delta_1}$ (constant plus power model) and $g(v, \delta) = \exp(\delta v)$ (exponential model). See [5, ch. 5] for a discussion of how to choose a variance function model.

We model the variability of errors in (1) as

$$\mathrm{var}(e_{ijk}|\mathbf{b}_i) = \sigma_j^2\, g_j^2(v_i, \boldsymbol{\delta}_j),\ v_i = h(\boldsymbol{\mu}_i),\ \boldsymbol{\mu}_i = \boldsymbol{\beta} + \mathbf{b}_i, \tag{3}$$

allowing each method to have its own variance function of the *common* covariate v. With the additional assumption of normality, the proposed heteroscedastic mixed-effects model for method comparison data is model (1) where

$$e_{ijk}|\mathbf{b}_i \sim \text{ independent } \mathcal{N}_1\big(0, \sigma_j^2\, g_j^2\big(v_i, \boldsymbol{\delta}_j\big)\big), \quad \mathbf{b}_i \sim \text{ independent } \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Psi}). \tag{4}$$

The homoscedastic model of Section 2.1 is a special case of this model when $\boldsymbol{\delta}_1 = \boldsymbol{\delta}_2 = 0$. Note that, unlike the homoscedastic case, the random effects and errors in (4) are not independent because the error variance function involves $\mathbf{b}_i$ through $\boldsymbol{\mu}_i$ in v_i.

Next, let $\mathbf{1}_n$ denote a $n \times 1$ vector of ones. Define $\mathbf{X}_i$ and $\mathbf{Z}_i$ as $n_i \times 2$ design matrices

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{1}_{n_{i1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_{i2}} \end{bmatrix}, \quad \mathbf{Z}_i = \mathbf{X}_i.$$

Also define $\boldsymbol{\Sigma}_{ij}(v_i)$ as a $n_{ij} \times n_{ij}$ diagonal matrix and $\boldsymbol{\Sigma}_i(v_i)$ as a $n_i \times n_i$ diagonal matrix,

$$\boldsymbol{\Sigma}_{ij}(v_i) = \mathrm{diag}\left\{\sigma_j^2\, g_j^2(v_i, \boldsymbol{\delta}_j), \ldots, \sigma_j^2\, g_j^2\big(v_i, \boldsymbol{\delta}_j\big)\right\}, \quad \boldsymbol{\Sigma}_i(v_i) = \mathrm{diag}\{\boldsymbol{\Sigma}_{i1}(v_i), \boldsymbol{\Sigma}_{i2}(v_i)\}.$$

We can now write the proposed heteroscedastic model in the matrix form as

$$\begin{aligned} &\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \quad \mathbf{e}_i|\mathbf{b}_i \sim \text{ independent } \mathcal{N}_{n_i}(\mathbf{0}, \boldsymbol{\Sigma}_i(v_i)), \\ &\mathbf{b}_i \sim \text{ independent } \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Psi}), \quad i = 1, \ldots, m, \end{aligned} \tag{5}$$

with $\mathbf{e}_i$ as the $n_i \times 1$ error vector. A hierarchical representation of this model is as follows:

$$\begin{aligned} &\begin{bmatrix} \mathbf{Y}_{i1} \\ \mathbf{Y}_{i2} \end{bmatrix} |(b_{i1}, b_{i2}) \sim \mathcal{N}_{n_i}\left(\begin{bmatrix} (\beta_1 + b_{i1})\mathbf{1}_{n_{i1}} \\ (\beta_2 + b_{i2})\mathbf{1}_{n_{i2}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{i1}(v_i) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{i2}(v_i) \end{bmatrix} \right), \\ &b_{i2}|b_{i1} \sim \mathcal{N}_1\big(\rho(\psi_2/\psi_1)b_{i1}, \psi_2^2(1-\rho^2)\big), \quad b_{i1} \sim \mathcal{N}_1(0, \psi_1^2). \end{aligned} \tag{6}$$

Here the model parameter vector $\boldsymbol{\theta}$ consists of $\boldsymbol{\beta}$, the elements in the upper triangle of $\boldsymbol{\Psi}$, and σ_1^2, σ_2^2, $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$.

The marginal density of $\mathbf{Y}_i$ under this model can be expressed as

$$f_{\boldsymbol{\theta}}(\mathbf{y}_i) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\boldsymbol{\theta}}(\mathbf{y}_i, \mathbf{b}_i)\, d\mathbf{b}_i = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\boldsymbol{\theta}}(\mathbf{y}_{i1}|\mathbf{b}_i)\, f_{\boldsymbol{\theta}}(\mathbf{y}_{i2}|\mathbf{b}_i)\, f_{\boldsymbol{\theta}}(\mathbf{b}_i)\, d\mathbf{b}_i, \tag{7}$$

where we have used the conditional independence of $\mathbf{Y}_{i1}$ and $\mathbf{Y}_{i2}$ given $\mathbf{b}_i$ to write $f_{\boldsymbol{\theta}}(\mathbf{y}_i|\mathbf{b}_i) = f_{\boldsymbol{\theta}}(\mathbf{y}_{i1}|\mathbf{b}_i)\, f_{\boldsymbol{\theta}}(\mathbf{y}_{i2}|\mathbf{b}_i)$. The densities involved in this two-dimensional integral are those of normal distributions and can be obtained from (6). The situation is somewhat simpler when v_i depends only on μ_{i1}, as may be the case when there is a reference method in the comparison. In this case, Proposition 1 in Appendix shows that b_{i2} can be explicitly integrated out to get the one-dimensional integral

$$f_{\boldsymbol{\theta}}(\mathbf{y}_i) = \int_{-\infty}^{\infty} f_{\boldsymbol{\theta}}(\mathbf{y}_i, b_{i1})\, db_{i1} = \int_{-\infty}^{\infty} f_{\boldsymbol{\theta}}(\mathbf{y}_{i1}|b_{i1})\, f_{\boldsymbol{\theta}}(\mathbf{y}_{i2}|b_{i1})\, f_{\boldsymbol{\theta}}(b_{i1})\, db_{i1}, \tag{8}$$

where the densities involved are those of normal distributions given in (A.2).

Unfortunately the marginal density $f_{\boldsymbol{\theta}}(\mathbf{y}_i)$, given by either (7) or (8), and hence the likelihood function,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{m} f_{\boldsymbol{\theta}}(\mathbf{y}_i),$$

is not available in a closed-form because $\mathbf{b}_i$ enters the model nonlinearly via the variance function, precluding it from being explicitly integrated out. A similar issue arises in nonlinear mixed-effects models [18, ch. 6] and generalized linear mixed-effects models [22, ch. 4-6].

2.3 Model fitting

The lack of a closed-form for the likelihood function causes difficulty in model fitting. We next discuss two approaches to deal with it — one computes the likelihood by numerically evaluating the integrals in (7) and (8), and the other approximates the model (5) so that the corresponding likelihood function is available in a closed-form. In either case, the model is fit by maximizing the resulting approximate likelihood function with respect to $\boldsymbol{\theta}$ to get $\hat{\boldsymbol{\theta}}$ as an approximate ML estimator of $\boldsymbol{\theta}$.

2.3.1 Approach 1: Numerical computation of likelihood

To treat the integrals in (7) and (8) in a unified manner, consider computing the integral $f_{\boldsymbol{\theta}}(\mathbf{y}_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{\theta}}(\mathbf{y}_i, \mathbf{u}_i)\, d\mathbf{u}_i$, where $\mathbf{u}_i$ is a $d \times 1$ vector. When the variance covariate v_i depends on both (μ_{i1}, μ_{i2}), $\mathbf{u}_i$ plays the role of $\mathbf{b}_i$ with $d = 2$ and the integral equals (7). On the other hand, when v_i depends only on μ_{i1}, $\mathbf{u}_i$ plays the role of b_{i1} with $d = 1$

and the integral equals (8). Let $\hat{\mathbf{u}}_i$ be a minimizer of the negative log-density, $l_{\boldsymbol{\theta}}(\mathbf{y}_i, \mathbf{u}_i) = -\log f_{\boldsymbol{\theta}}(\mathbf{y}_i, \mathbf{u}_i)$, with respect to $\mathbf{u}_i$. Also let

$$\mathbf{l}''_{\boldsymbol{\theta}}(\mathbf{y}_i, \hat{\mathbf{u}}_i) = \left.\frac{\partial^2 l_{\boldsymbol{\theta}}(\mathbf{y}_i, \mathbf{u}_i)}{\partial \mathbf{u}_i \partial \mathbf{u}_i^T}\right|_{\mathbf{u}_i = \hat{\mathbf{u}}_i}$$

be the corresponding $d \times d$ Hessian matrix evaluated at $\hat{\mathbf{u}}_i$.

A simple method for computing the integral is Laplace approximation [23], which gives

$$f_{\boldsymbol{\theta}}(\mathbf{y}_i) \approx (2\pi)^{d/2} |\mathbf{l}''_{\boldsymbol{\theta}}(\mathbf{y}_i, \hat{\mathbf{u}}_i)|^{-1/2} f_{\boldsymbol{\theta}}(\mathbf{y}_i, \hat{\mathbf{u}}_i). \tag{9}$$

We refer to this method as the “LA-L” (likelihood approximation-Laplace) method. Another alternative is a Gauss-Hermite quadrature method [23]. To describe it, let $z_1, \ldots, z_M$ be the nodes and $w_1, \ldots, w_M$ be the associated weights for one-dimensional quadrature with kernel $\exp(-z^2)$. The nodes are roots of the Mth degree Hermite polynomial. The quadrature grid in the d-dimensional space is $(z_1, \ldots, z_M) \times \ldots \times (z_1, \ldots, z_M)$. Let the $d \times 1$ vector $\mathbf{z}_r = (z_{r_1}, \ldots, z_{r_d})$ denote a node in this grid. The total number of such nodes is M^d. For greater accuracy of approximation, these nodes need to be centered and scaled [24, 25] as

$$\mathbf{a}_{i,r} = \hat{\mathbf{u}}_i + 2^{1/2}\, \mathbf{l}''_{\boldsymbol{\theta}}(\mathbf{y}_i, \hat{\mathbf{u}}_i)^{-1/2}\, \mathbf{z}_r.$$

The approximated integral in this case is [26]

$$f_{\boldsymbol{\theta}}(\mathbf{y}_i) \approx 2^{d/2}\, |\mathbf{l}''_{\boldsymbol{\theta}}(\mathbf{y}_i, \hat{\mathbf{u}}_i)|^{-1/2} \sum_{r_1=1}^{M} \cdots \sum_{r_d=1}^{M} f_{\boldsymbol{\theta}}(\mathbf{y}_i, \mathbf{a}_{i,r}) \prod_{s=1}^{d} w_{r_s} \exp(z_{r_s}^2). \tag{10}$$

We refer to this method as the “LA-Q” (likelihood approximation-quadrature) method.

Note that the Laplace approximation is a special case of the quadrature method when $M = 1$. In this case, the sole node $z_1 = 0$ has weight $w_1 = \pi^{1/2}$, implying $\mathbf{a}_{i,r} = \hat{\mathbf{u}}_i$, and (10) reduces to (9). Both the accuracy of the quadrature method and its computational burden increase with M. Thus, the LA-Q method with $M > 1$ is not only more accurate but also more computationally demanding than the LA-L method. In practice, 20-30 nodes generally provide acceptable accuracy for the quadrature method. Also, the arguments in [27] can be used to show that the accuracy of the LA-L method tends to increase as n_i, the number of measurements on ith subject, increases. In either case, the resulting approximate likelihood function can be numerically maximized, e.g., using the `optim` function in `R`, to get $\hat{\boldsymbol{\theta}}$.

2.3.2 Approach 2: Model approximation

This approach follows [18, ch. 6] and approximates the model (5) by replacing the unobservable true value $\boldsymbol{\mu}_i$ in the variance function (3) with an *observable* quantity, $\boldsymbol{\mu}_i^* = (\mu_{i1}^*, \mu_{i2}^*)$. This $\boldsymbol{\mu}_i^*$ is expected to be close to $\boldsymbol{\mu}_i$ but is free of $\mathbf{b}_i$ and is held *fixed* during model fitting. It leads to $v_i^* = h(\boldsymbol{\mu}_i^*)$ as the approximate variance covariate and

$$\mathrm{var}(e_{ijk}) \approx \sigma_j^2\, g_j^2\big(v_i^*, \boldsymbol{\delta}_j\big) \tag{11}$$

as the approximate variance function. By analogy with (5), this approximate heteroscedastic mixed-effects model can be written as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_i\mathbf{b}_i+\mathbf{e}_i,\ \mathbf{b}_i \sim \text{ independent } \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Psi}),\ \mathbf{e}_i \sim \text{ independent } \mathcal{N}_{n_i}(\mathbf{0}, \boldsymbol{\Sigma}_i(v_i^*)), \tag{12}$$

$i = 1, \ldots, m$. In this formulation, $\mathbf{b}_i$ and $\mathbf{e}_i$ are independent, and marginally

$$\mathbf{Y}_i \sim \text{ independent } \mathcal{N}_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Pi}_i(v_i^*)),\quad \boldsymbol{\Pi}_i(v_i^*) = \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^T + \boldsymbol{\Sigma}_i(v_i^*).$$

We now discuss two natural choices for $\boldsymbol{\mu}_i^*$ and the fitting of the resulting model (12). The first is a standard choice recommended by [5, ch. 5] and [18, ch. 6]. It takes the best linear unbiased predictor (BLUP) of $\boldsymbol{\mu}_i$ under (12) as $\boldsymbol{\mu}_i^*$, which can be written as $\boldsymbol{\mu}_{i,\text{blup}} = E(\boldsymbol{\mu}_i|\mathbf{Y}_i) = \boldsymbol{\beta} + \mathbf{b}_{i,\text{blup}}$, with

$$\mathbf{b}_{i,\text{blup}} = E(\mathbf{b}_i|\mathbf{Y}_i) = \boldsymbol{\Psi}\mathbf{Z}_i^T\{\boldsymbol{\Pi}_i(v_i^*)\}^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})$$

denoting the BLUP of $\mathbf{b}_i$ [28, ch. 4]. Notice that this $\boldsymbol{\mu}_{i,\text{blup}}$, which is needed to compute the covariate v_i^* for model fitting, itself depends on the unknown $\boldsymbol{\theta}$. This calls for an iteratively reweighted scheme to fit (12). This scheme repeats the following two steps in each iteration until convergence [5, ch. 5]: In iteration $t = 1, 2, \ldots$,

(a) use the current $\boldsymbol{\theta}^{(t)}$ to compute the BLUP $\boldsymbol{\mu}_{i,\text{blup}}^{(t)}$ and get $v_i^{*(t)} = h(\boldsymbol{\mu}_i^{*(t)})$;

(b) hold $v_i^{*(t)}$ fixed and maximize the likelihood function of model (12) with $\boldsymbol{\Sigma}_i(v_i^{*(t)})$ as the error variance matrix to produce an updated estimate $\boldsymbol{\theta}^{(t+1)}$.

Any existing software for fitting mixed-effects models can be used for the maximization step and the estimates from the homoscedastic fit can be used as the starting point $\boldsymbol{\theta}^{(1)}$. The convergence, although not guaranteed, can be monitored by examining the maximum relative change in components of $\boldsymbol{\theta}$ or the relative change in likelihood in two successive iterations. Nevertheless, in practice, one does not need to wait till convergence because, as in heteroscedastic regression models [19], 2-3 iterations are generally enough to get a good estimate of $\boldsymbol{\theta}$. The model fit in the last iteration is used for all subsequent inference. We refer to this estimation method as the "MA-B" (model approximation-BLUP) method.

The second choice for $\boldsymbol{\mu}_i^*$ is to take $\boldsymbol{\mu}_i^* = \overline{\mathbf{Y}}_i$, the vector of sample means. In this case, the model (12), with $\boldsymbol{\mu}_i^*$ held fixed, can be fit via ML using any mixed-effects models software. We refer to this method as the "MA-M" (model approximation-mean) method. Needless to say, the model fitting in this case is much simpler than before. Note that both choices for $\boldsymbol{\mu}_i^*$ — $\boldsymbol{\mu}_{i,\text{blup}}$ and $\overline{\mathbf{Y}}_i$ — are linear unbiased predictors of $\boldsymbol{\mu}_i$ in that $E(\boldsymbol{\mu}_{i,\text{blup}} - \boldsymbol{\mu}_i) = \mathbf{0} = E(\overline{\mathbf{Y}}_i - \boldsymbol{\mu}_i)$. When $\boldsymbol{\theta}$ is known, $\boldsymbol{\mu}_{i,\text{blup}}$ is obviously a better predictor than $\overline{\mathbf{Y}}_i$ due to its optimality property. But the optimality is not guaranteed when $\boldsymbol{\theta}$ is estimated.

2.4 Inference on model parameters

The previous section discussed four methods for computing approximate ML estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. For further inference, we proceed as in nonlinear mixed-effects models [18, ch. 6] and generalized linear mixed-effects models [22, ch. 4-6] by essentially ignoring the fact that we are working with either the approximate likelihood or the approximate model. Thus, regardless of the estimation method used, when m is large, we can approximate the distribution of $\hat{\boldsymbol{\theta}}$

by a normal distribution with mean $\boldsymbol{\theta}$ and variance $\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}$, where

$$\mathbf{I}(\hat{\boldsymbol{\theta}}) = -\left.\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

is the observed information matrix. Note that $L(\boldsymbol{\theta})$ here denotes the likelihood function actually maximized to get $\hat{\boldsymbol{\theta}}$. The theoretical justification for this result is easiest to give in case of quadrature approximation (10) of likelihood. When the number of nodes M is large, the error in approximation of the integral can be effectively ignored, and the asymptotic normality follows from the standard large sample theory of ML estimators [29]. One can get this result for Laplace approximation (9) of likelihood as well by adapting [27] assuming that $\min_{i=1}^{m}\{n_i\}$ is also large in addition to m.

The approximate normality of $\hat{\boldsymbol{\theta}}$ can be used in the usual manner for likelihood-based inference on $\boldsymbol{\theta}$. The derivatives needed for $\mathbf{I}$ can be computed numerically, e.g., using the `numDeriv` package [30] in `R`. In applications, we are particularly interested in testing the null hypothesis of homoscedasticity, $H_0 : \boldsymbol{\delta}_1 = \mathbf{0} = \boldsymbol{\delta}_2$. We can use either a likelihood ratio test or a score test for this purpose. The likelihood ratio statistic is twice the difference in the negative log-likelihoods of the "full" heteroscedastic model and the "reduced" homoscedastic model, which assumes H_0 to be true. To define the score statistic, let

$$\mathbf{s}(\boldsymbol{\theta}) = \frac{\partial \log L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$$

be the score function. Write $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1 = (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$ and $\boldsymbol{\theta}_2$ is the vector of remaining model parameters. Thus, the null hypothesis is $H_0 : \boldsymbol{\theta}_1 = \mathbf{0}$. Next, conformably partition $\mathbf{s} = (\mathbf{s}_1, \mathbf{s}_2)$ and

$$\mathbf{I} = \begin{bmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{21} & \mathbf{I}_{22} \end{bmatrix}.$$

Let $\hat{\boldsymbol{\theta}}_{2,0}$ be the ML estimator of $\boldsymbol{\theta}_2$ under the reduced model. Define $\hat{\boldsymbol{\theta}}_0 = (\boldsymbol{\theta}_1 = \mathbf{0}, \hat{\boldsymbol{\theta}}_{2,0})$. Then the score statistic [31] for testing H_0 is $\mathbf{s}(\hat{\boldsymbol{\theta}}_0)^T\, \mathbf{I}_{11.2}(\hat{\boldsymbol{\theta}}_0)^{-1}\, \mathbf{s}(\hat{\boldsymbol{\theta}}_0)$, where $\mathbf{s}(\hat{\boldsymbol{\theta}}_0)$ represents the score function evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_0$ and $\mathbf{I}_{11.2} = \mathbf{I}_{11} - \mathbf{I}_{12}\mathbf{I}_{22}^{-1}\mathbf{I}_{12}^T$. Notice that the ML estimator of $\boldsymbol{\theta}$ under the full model is not needed to compute the score statistic. When m is large, the null distributions of both likelihood ratio and score statistics can be approximated by a chi-squared distribution with degrees of freedom equal to the number of elements in $(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$. Thus, the p-values for testing H_0 can be obtained by finding probabilities under this distribution to the right of the observed values of the test statistics.

3 A simulation study

In this section, we use Monte Carlo simulation to evaluate and compare finite sample performance of the four model fitting methods proposed in Section 2.3. Two of these methods — LA-L and LA-Q — are based on likelihood approximation, whereas the other two — MA-B and MA-M — are based on model approximation. Of specific interest are the accuracies of the point and interval estimators of model parameters and the test of homoscedasticity. To measure accuracy, we focus on bias and mean squared error (MSE) for a point estimator, coverage probability for an interval estimator, and type I error probability for a test.

The data are simulated on $m = 50$ subjects from the true model (5) using a balanced design with $n_{ij} = 2$ and 3 replications per method. The power model, $g_j(v, \delta_j) = |v|^{\delta_j}$, $j = 1, 2$, with v as the true value of method 1, is used as the variance function model. We let $\delta_1 = \delta_2 = \delta$ in the simulations. Table 1 summarizes the actual parameter values that we use, which are motivated by the estimates for the cholesterol data in Section 5. They encompass a variety of scenarios we have seen in applications, including measurement methods with characteristics that range from "quite different" to "very similar" and a level of heteroscedasticity that ranges from "none" to "quite substantial." Besides the error variances here are kept small in relation to the between-subject variance as this is normally the case in method comparison studies.

After simulating a data set, we compute point and interval estimates of model parameters and carry out the test of homoscedasticity using all four fitting methods. The density given by (8) is used for LA-L and LA-Q methods. The latter uses $M = 30$ nodes. Also, only three iterations of the iteratively reweighted scheme are employed in case of the MA-B method. Further, to improve finite sample performance, inference on variance parameters and correlation is performed after applying a normalizing transformation. In particular, a log transformation is applied to σ_1^2, σ_2^2, ψ_1^2 and ψ_2^2, and the Fisher's z-transformation, $z(\rho) = \tanh^{-1}(\rho)$, is applied to ρ. The process of simulating data and performing inference is repeated 500 times, and the following estimates are computed: biases and MSE's of point estimators, coverage probabilities of confidence intervals with 95% nominal level, and type I error probabilities of tests of homoscedasticity with 5% nominal level. Here we only present results for $n_{ij} = 2$ and $\delta = 0, 1.1$; those for $n_{ij} = 3$ or $\delta = 0.9, 1$ are omitted as they lead to essentially the same qualitative conclusions that we describe below.

Tables 2 and 3 present estimated biases and MSE's of the four point estimators when $\delta = 0$ and $\delta = 1.1$, respectively. In both cases, the biases for (β_1, β_2) are negligible relative to their true values. The biases are negative for $(\log \psi_1^2, \log \psi_2^2)$ and positive for $z(\rho)$, albeit they are small. The conclusion, however, is less clear-cut for $(\log \sigma_1^2, \log \sigma_2^2)$ when the results for $\delta = 0.9, 1$ are included (not presented). The biases in these cases are also small but they may be positive as well as negative. We also see that the MA-B and MA-M estimators appear equally accurate as they have nearly identical biases and MSE's. While some differences exist in the accuracies of the LA-L and LA-Q estimators, the differences are not substantial. There is also evidence that the LA-Q estimators may be slightly superior to the other three in terms of MSE, especially for (β_1, β_2). But on the whole, there is little practical difference in the accuracies of the four estimators.

Table 4 shows estimated coverage probabilities of 95% confidence intervals computed using the four fitting methods. The entries for MA-M and MA-B methods are practically the same and they can be considered reasonably close to 95%. However, the same cannot be said for the LA-L and LA-Q methods; most entries are quite less than 95% and some even fall below 90%. This comparison shows that the methods based on model approximation are superior to those based on likelihood approximation for constructing confidence intervals.

Table 5 reports estimated type I error probabilities for 5% level likelihood ratio test and score test for homoscedasticity using the four fitting methods. For the likelihood ratio test, all four methods seem equally accurate, although the test may be a bit conservative, especially in case of setting 3. For the score test, the LA-L method does not work well as its type I error probability is substantially less than 5%. The other three methods work well, but they may be a bit liberal. We also examined the normal quantile-quantile plots of parameter estimates from all four methods (not presented here) and found no obvious departure from normality in any case.

Overall, these results suggest that the two estimation methods based on model approximation are equally accurate and they lead to confidence intervals and tests that have reasonably good performance with $m = 50$ subjects. While in some instances, the MSE's of these estimators may be slightly greater than those based on likelihood approximation, the differences are not large enough to be practically important. Besides the former methods tend to produce more accurate confidence intervals than the latter methods. Next, recall that the methods based on model approximation are easier to implement than those based on likelihood approximation. In particular, the MA-M method is easiest to apply; it can be implemented using any software for fitting mixed-effects models. Thus, there appears little reason to prefer any other method over the MA-M method.

Upon repeating the simulations with $m = 30$ subjects, we see that the findings regarding relative merits of the estimation methods continue to hold, but the tests and confidence interval became somewhat less accurate. In particular, the coverage probabilities of 95% intervals are about 93-94% and the type I error probabilities of 5% level likelihood ratio test of homoscedasticity is about 6-7%. Thus, it appears that data from about 50 subjects are necessary for the tests and confidence intervals to be reasonably accurate. The number of replications, provided it is at least two, does not have a noteworthy effect on this accuracy.

4 Evaluation of agreement under the heteroscedastic model

In this section, we discuss how to evaluate agreement in two measurement methods assuming the heteroscedastic model (5) for the method comparison data. We proceed as in [11, 21] to adapt the methodology developed assuming a homoscedastic model.

Let $(\tilde{Y}_1, \tilde{Y}_2)$ denote the paired measurements by the two methods on a randomly selected subject from the population. We can think of $(\tilde{Y}_1, \tilde{Y}_2)$ as a "typical" measurement pair. Under (5), the marginal distribution of $(\tilde{Y}_1, \tilde{Y}_2)$ for a *fixed* (known) value v_0 of the variance covariate v is,

$$\begin{bmatrix} \tilde{Y}_1 \\ \tilde{Y}_2 \end{bmatrix} \sim \mathcal{N}_2 \left(\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \begin{bmatrix} \psi_1^2 + \sigma_1^2 \, g_1^2(v_0, \boldsymbol{\delta}_1) & \psi_{12} \\ \psi_{12} & \psi_2^2 + \sigma_2^2 \, g_2^2(v_0, \boldsymbol{\delta}_2) \end{bmatrix} \right), \quad v_0 \in \mathcal{V}_0, \tag{13}$$

where $\mathcal{V}_0$ is the range of values of v_0 of interest. In practice, $\mathcal{V}_0$ can be taken as the observed measurement range.

4.1 Inference on an agreement measure

Let ϕ denote any scalar measure of agreement between the two methods [1, 2]. This ϕ is a known function of the model parameter vector $\boldsymbol{\theta}$ quantifying how concentrated the bivariate distribution of $(\tilde{Y}_1, \tilde{Y}_2)$ is around the 45^o line. In the heteroscedastic case, this distribution, given by (13), depends on v_0. Therefore, ϕ is a function of v_0 as well as $\boldsymbol{\theta}$. Let this function be denoted by a, i.e., $\phi(v_0) = a(\boldsymbol{\theta}, v_0)$, where a has a known form depending on the agreement measure being considered. A natural estimator for $\phi(v_0)$ is

$\hat{\phi}(v_0) = a(\hat{\boldsymbol{\theta}}, v_0)$, which is obtained by simply plugging-in $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ in a. Since for a large m, $\hat{\boldsymbol{\theta}}$ is approximately normal with mean $\boldsymbol{\theta}$ and variance $\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}$, it follows from the delta method [29] that $\hat{\phi}(v_0)$ also approximately follows a $\mathcal{N}_1(\phi(v_0), \mathbf{G}^T(v_0)\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{G}(v_0))$ distribution, with $\mathbf{G}(v_0) = \partial a(\boldsymbol{\theta}, v_0)/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ denoting the gradient vector. The derivative needed here can be obtained in a closed-form (see, e.g., [11]) or it also be computed numerically. This result can be used for inference on the agreement measure ϕ.

In particular, if ϕ is such that small values for it imply good agreement, we can compute an approximate $100(1-\alpha)\%$ pointwise *upper* confidence band for $\phi(v_0)$ as:

$$\hat{\phi}(v_0) + z_{1-\alpha}\{\mathbf{G}^T(v_0)\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{G}(v_0)\}^{1/2}, \quad v_0 \in \mathcal{V}_0,$$

where z_α is the αth quantile of a $\mathcal{N}_1(0,1)$ distribution. Similarly, if ϕ is such that large values for it imply good agreement, we can compute an approximate $100(1-\alpha)\%$ pointwise *lower* confidence band for $\phi(v_0)$ as:

$$\hat{\phi}(v_0) - z_{1-\alpha}\{\mathbf{G}^T(v_0)\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{G}(v_0)\}^{1/2}, \quad v_0 \in \mathcal{V}_0.$$

These one-sided confidence bands can be used to find subsets of the measurement range, if any, where the methods have sufficient agreement for interchangeable use. To improve finite sample accuracy, these bands may be computed after applying a suitable normalizing transformation to ϕ and transforming the results back to the original scale. In absence of heteroscedasticity, this methodology reduces to the one considered in [21] for two methods.

4.2 Heteroscedastic versions of two common agreement measures

The heteroscedastic version of any agreement measure ϕ can be obtained by using the distribution (13) in its definition. Below we provide expressions for two popular measures, *concordance correlation coefficient* [CCC; 32] and *total deviation index* [TDI; 33–36].

The CCC measure, introduced by [32], is defined as

$$\text{CCC} = \frac{2\, cov(\tilde{Y}_1, \tilde{Y}_2)}{(E(\tilde{Y}_1) - E(\tilde{Y}_2))^2 + var(\tilde{Y}_1) + var(\tilde{Y}_2)}.$$

It lies between -1 and +1 and a large value for it implies good agreement. Its heteroscedastic version is obtained by using (13) for the moments of $(\tilde{Y}_1, \tilde{Y}_2)$, leading to

$$\text{CCC } (v_0) = \frac{2\psi_{12}}{(\beta_1 - \beta_2)^2 + \psi_1^2 + \sigma_1^2\, g_1^2(v_0, \delta_1) + \psi_2^2 + \sigma_2^2\, g_2^2(v_0, \delta_2)}.$$

Next, the TDI measure, introduced by [33], is based on the difference $\tilde{D} = \tilde{Y}_1 - \tilde{Y}_2$. It is defined as the p_0th percentile of $|\tilde{D}|$ for a given large probability p_0. For normally distributed $\tilde{D}$, it can be expressed as

$$\text{TDI } (p_0) = sd(\tilde{D}) \left\{ \chi_1^2 \left(p_0, \{E(\tilde{D})/sd(\tilde{D})\}^2 \right) \right\}^{1/2},$$

where $\chi_1^2(p_0, \Delta)$ is the p_0th quantile of a chi-square distribution with one degree of freedom and non-centrality parameter Δ. This measure is non-negative and a small value for it indicates good agreement. To define its heteroscedastic version, note from (13) that for a given v_0, $\tilde{D}$ follows a $\mathcal{N}_1(\beta_1 - \beta_2, \tau^2(v_0))$ distribution where

$$\tau^2(v_0) = \psi_1^2 + \psi_2^2 - 2\psi_{12} + \sigma_1^2\, g_1^2(v_0, \delta_1) + \sigma_2^2\, g_2^2(v_0, \delta_2).$$

Therefore,

$$\text{TDI } (v_0, p_0) = \tau(v_0) \left\{ \chi_1^2 \left(p_0, \{(\beta_1 - \beta_2)/\tau(v_0)\}^2 \right) \right\}^{1/2}.$$

The inference on these measures proceeds as described in the previous subsection. For greater accuracy, it is common to use the Fisher's z-transformation of CCC and the log transformation of TDI [34]. The other measures in [1] can be handled in a similar manner.

5 Analysis of cholesterol data

We now return to the cholesterol data introduced in Section 1 where we saw a clear evidence of magnitude-dependent heteroscedasticity. Our first task is to find an adequate model for these data. Since here the Cobas Bio assay (method 1) serves as a reference method, we take its true measurement μ_{i1} as the variance covariate v_i. Further, on the basis of a preliminary analysis of residuals in Figure 2, we take the power model, $g_j(v, \delta_j) = |v|^{\delta_j}$, $j = 1, 2$, as the variance function model for the assays. The heteroscedastic model (5) is fit using all four estimation methods described in Section 2. Table 6 presents the resulting estimates and their standard errors for the nine model parameters. Remarkably both likelihood approximation methods and both model approximation methods produce identical results when rounded to two decimal places. Further, there is no substantive difference in the two sets of estimates. This is not surprising given the findings of the simulation study. Therefore, hereafter we present results only for the MA-M method, which is the simplest one to implement.

The p-values for both the likelihood ratio test and the score test of homoscedasticity is practically zero, confirming that the error variances are nonconstant. Figure 3 shows residual plots for the fitted heteroscedastic model. As there is no discernible pattern in these plots, we can conclude that the proposed model fits well to the data.

Our next task is to use the fitted model to evaluate agreement between the two assays over the observed measurement range of 45 mg/dl to 370 mg/dl. The estimates in Table 6 confirm that the Ektachem measurements have a larger mean and a larger error variation compared to Cobas Bio. Their between-subject variations, however, are quite comparable. Upon using these estimates in (13), we can see that the error standard deviation increases monotonically from 0.43 to 3.71 in case of Cobas Bio and from 0.59 to 4.72 in case of Ektachem. Their correlation remains very high — above 0.99 — over the entire measurement range. The mean of the difference between their typical measurements is about 5.5 mg/dl and its standard deviation increases monotonically from 7.32 to 9.45. Next, we apply the procedure described in Section 4 for inference on CCC and TDI (with $p = 0.90$). Their 95% pointwise one-sided confidence bands — lower band for CCC and upper band for TDI — are computed by first applying the Fisher's z-transformation of CCC and the log transformation of TDI. The results as functions of the cholesterol level are presented in Figure 4.

We see that the lower confidence bound for CCC decreases and the upper confidence bound for TDI increases as the cholesterol level increases. Thus, in an absolute sense, the extent of agreement between the assays becomes progressively worse, albeit only by a small amount, with increasing magnitude. The CCC lower bounds are quite close to one, suggesting excellent agreement between the assays over the entire measurement range. But this conclusion is misleading because the between-subject variation in these data is much greater than the within-subject variation (see Figure 1), guaranteeing a large CCC regardless of the true extent of agreement. This drawback of CCC is well-known in the literature [37, 38].

A better picture of agreement is given by the TDI whose upper bound increases from 17 to 19.5 mg/dl as the cholesterol level increases from 45 to 370 mg/dl. The value of 17, e.g., shows that 90% of differences in measurements from the assays fall within ±17 when the true value is 45. Such a difference is unacceptably large relative to the true value. On the other hand, a difference as large as ±19.5 may be acceptable when the true value is 370. Thus, we may conclude satisfactory agreement between the assays for large cholesterol values but not for small values. It may be noted that the 95% confidence bound for TDI when the homoscedastic model of Section 2.1 is fit is 17.6, which does not depend on the cholesterol value. While 17.6 is not terribly far from either 17 or 19.5, clearly the extent of agreement on the basis of homoscedastic model is underestimated for small cholesterol values and is overestimated for large cholesterol values.

We also repeat the analysis based on heteroscedastic model after subtracting 5.5 from Ektachem measurements so that the two assays have comparable means. The TDI bounds now range between 12 to 15, showing improvement in the extent of agreement. These conclusions remain unchanged when the analysis is repeated by taking the average true measurement, i.e., $v_i = (\mu_{i1} + \mu_{i2})/2$, as the variance covariate.

6 Discussion

In this article, we use a basic heteroscedastic mixed-effects model for modeling method comparison data that allows the variability of measurements to depend on the magnitude. The fitted model is then used to develop a methodology for agreement evaluation in two measurement methods. Our approach is flexible in that it can accommodate balanced or unbalanced data designs and it works with any scalar measure of agreement. It can also be extended to handle more than two measurement methods and incorporate covariates in a straightforward manner by following [11, 21].

A potential limitation of our approach is the assumption that the measurements are replicated. Although it is a good idea to have replications [11, 13], it is common in applications to have only unreplicated data. In principle, our methodology can be applied to unreplicated data as well by suitably modifying the model, but it often does not work in practice because there may not be enough information in the data to estimate all model parameters. It may be possible to deal with this issue by fitting the model using a Bayesian approach with informative prior distributions, but further research is needed for this investigation.

Acknowledgements

The authors thank two anonymous reviewers and the Associate Editor for their comments that greatly improved this article. They are also grateful to Professor Vernon Chinchilli for providing the cholesterol data, and to the Texas Advanced Computing Center at The University of Texas at Austin for providing HPC resources for conducting the simulation studies reported in this paper.

References

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

2. Choudhary PK. Interrater agreement 2009; In *Methods and Applications of Statistics in the Life and Health Sciences*, pp. 461-480, Balakrishnan, N. et al. (Editors), John Wiley: New York.

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

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

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

6. Carrasco JL, Jover L. Estimating the generalized concordance correlation coefficient through variance components. *Biometrics* 2003; **59**:849–858.

7. Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. *Journal of Biopharmaceutical Statistics* 2007; **17**:571–582.

8. Carstensen B, Simpson J, Gurrin LC. Statistical models for assessing agreement in method comparison studies with replicate measurements. *The International Journal of Biostatistics* 2008; **4**:article 16.

9. 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.

10. Roy A. An application of linear mixed effects model to assess the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics* 2009; **19**:150–173.

11. Choudhary PK. A tolerance interval approach for assessment of agreement in method comparison studies with repeated measurements. *Journal of Statistical Planning and Inference* 2008; **138**:1102–1115.

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

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

14. Hawkins DM. Diagnostics for conformity of paired quantitative measurements. *Statistics in Medicine* 2002; **21**:1913–1935.

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

16. Pinheiro JC, Bates D, DebRoy S, Sarkar D, R Development Core Team. *nlme: Linear and nonlinear mixed effects models*, 2012. `http://CRAN.R-project.org/package=nlme`.

17. R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2012. `http://www.R-project.org`.

18. Davidian M, Giltinan DM. *Nonlinear Models for Repeated Measurement Data.* Chapman & Hall/CRC Press: Boca Raton, FL, 1995.

19. Carroll RJ, Ruppert D. *Transformation and Weighting in Regression.* Chapman & Hall: New York, 1988.

20. Choudhary PK, Ng HKT. A tolerance interval approach for assessment of agreement using regression models for mean and variance. *Biometrics* 2006; **62**:288–296.

21. Choudhary PK, Yin K. Bayesian and frequentist methodologies for analyzing method comparison studies with multiple methods. *Statistics in Biopharmaceutical Research* 2010; **2**:122–132.

22. Stroup WW. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CRC, Boca Raton, FL, 2012.

23. Lange K. *Numerical Analysis for Statisticians*. 2nd edn. Springer, New York, 2010.

24. Liu Q, Pierce DA. A note on Gauss-Hermite quadrature. *Biometrika* 1994; **81**:624–629.

25. Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics* 1995; **4**:12–35.

26. Tuerlinckx F, Rijmen F, Verbeke G, De Boeck P. Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology* 2006; **59**:225–255.

27. Vonesh EF. A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika* 1996; **83**:447–452.

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

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

30. Gilbert P. *numDeriv: Accurate numerical derivatives*, 2011. `http://CRAN.R-project.org/package=numDeriv`.

31. Tarone RE. Score statistics 1988; In *Encyclopedia of Statistical Sciences*, pp. 304-308, Kotz, S., Johnson, N. L. and Read, C. B. (Editors), John Wiley: New York.

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

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

34. 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.

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

36. Escaramis G, Ascaso C, Carrasco JL. The total deviation index estimated by tolerance intervals to evaluate the concordance of measurement devices. *BMC Medical Research Methodology* 2010; **10**:article 31.

37. Atkinson G, Nevill A. Comment on the use of concordance correlation to assess the agreement between two variables. *Biometrics* 1997; **53**:775–777.

38. Barnhart HX, Lokhnygina Y, Kosinski AS, Haber MJ. Comparison of concordance correlation coefficient and coefficient of individual agreement in assessing agreement. *Journal of Biopharmaceutical Statistics* 2007; **17**:721–738.

Appendix A Technical details

Proposition 1. *Consider the heteroscedastic mixed-effects model* (5) *where the variance covariate v_i depends on vector $\boldsymbol{\mu}_i$ only through its first component μ_{i1}. Let $\mathbf{J}_n$ denote a $n \times n$*

matrix of ones. Then, the marginal density of $(\mathbf{Y}_{i1}, \mathbf{Y}_{i2}, b_{i1})$ *is*

$$f_{\boldsymbol{\theta}}(\mathbf{y}_{i1}, \mathbf{y}_{i2}, b_{i1}) = f_{\boldsymbol{\theta}}(\mathbf{y}_{i1}|\, b_{i1})\, f_{\boldsymbol{\theta}}(\mathbf{y}_{i2}|\, b_{i1})\, f_{\boldsymbol{\theta}}(b_{i1}), \quad i = 1, \dots m, \tag{A.1}$$

where

$$\begin{aligned} &\mathbf{Y}_{i1}|b_{i1} \sim \mathcal{N}_{n_{i1}}\big((\beta_1 + b_{i1})\mathbf{1}_{n_{i1}}, \boldsymbol{\Sigma}_{i1}(v_i)\big), \\ &\mathbf{Y}_{i2}|b_{i1} \sim \mathcal{N}_{n_{i2}}\big((\beta_2 + \rho(\psi_2/\psi_1)b_{i1})\mathbf{1}_{n_{i2}}, \boldsymbol{\Sigma}_{i2}(v_i) + \psi_2^2(1-\rho^2)\, \mathbf{J}_{n_{i2}}\big), \quad b_{i1} \sim \mathcal{N}_1(0, \psi_1^2). \end{aligned} \tag{A.2}$$

Proof. We can write the joint density of $(\mathbf{Y}_{i1}, \mathbf{Y}_{i2}, b_{i1}, b_{i2})$ as

$$f_{\boldsymbol{\theta}}(\mathbf{y}_{i1}, \mathbf{y}_{i2}, b_{i1}, b_{i2}) = f_{\boldsymbol{\theta}}(\mathbf{y}_{i1}, \mathbf{y}_{i2}|b_{i1}, b_{i2}) f_{\boldsymbol{\theta}}(b_{i2}|b_{i1}) f_{\boldsymbol{\theta}}(b_{i1}). \tag{A.3}$$

From conditional independence, the first term on the right is $f_{\boldsymbol{\theta}}(\mathbf{y}_{i1}|b_{i1}, b_{i2}) f_{\boldsymbol{\theta}}(\mathbf{y}_{i2}|b_{i1}, b_{i2})$. Next, under the assumptions, v_i is free of b_{i2}, implying that $f_{\boldsymbol{\theta}}(\mathbf{y}_{i1}|b_{i1}, b_{i2}) = f_{\boldsymbol{\theta}}(\mathbf{y}_{i1}|b_{i1})$. Therefore, the joint density in (A.3) can be written as

$$\begin{aligned} f_{\boldsymbol{\theta}}(\mathbf{y}_{i1}, \mathbf{y}_{i2}, b_{i1}, b_{i2}) &= f_{\boldsymbol{\theta}}(\mathbf{y}_{i1}|b_{i1}) f_{\boldsymbol{\theta}}(\mathbf{y}_{i2}|b_{i1}, b_{i2}) f_{\boldsymbol{\theta}}(b_{i2}|b_{i1}) f_{\boldsymbol{\theta}}(b_{i1}) \\ &= f_{\boldsymbol{\theta}}(\mathbf{y}_{i1}|b_{i1}) f_{\boldsymbol{\theta}}(\mathbf{y}_{i2}, b_{i2}|b_{i1}) f_{\boldsymbol{\theta}}(b_{i1}). \end{aligned}$$

Integrating out b_{i2} from this density gives the expression in (A.1). Now it just remains to identify the distributions in (A.2). The normal distributions of $\mathbf{Y}_{i1}|b_{i1}$ and b_{i1} follow directly from the hierarchical representation (6). Also from this representation,

$$\mathbf{Y}_{i2}|(b_{i1}, b_{i2}) \sim \mathcal{N}_{n_{i2}}\big((\beta_2 + b_{i2})\mathbf{1}_{n_{i2}}, \boldsymbol{\Sigma}_{i2}(v_i)\big), \quad b_{i2}|b_{i1} \sim \mathcal{N}_1\big(\rho(\psi_2/\psi_1)b_{i1}, \psi_2^2(1-\rho^2)\big).$$

Since b_{i2} appears linearly in the mean of $\mathbf{Y}_{i2}|(b_{i1}, b_{i2})$ and its variance is free of b_{i2}, we can marginalize over b_{i2} to see that $\mathbf{Y}_{i2}|b_{i1}$ has a normal distribution with

$$\begin{aligned} &E(\mathbf{Y}_{i2}|b_{i1}) = E(E(\mathbf{Y}_{i2}|b_{i1}, b_{i2})) = (\beta_2 + \rho(\psi_2/\psi_1)b_{i1})\mathbf{1}_{n_{i2}} \\ &\mathrm{var}(\mathbf{Y}_{i2}|b_{i1}) = E(\mathrm{var}(\mathbf{Y}_{i2}|b_{i1}, b_{i2})) + \mathrm{var}(E(\mathbf{Y}_{i2}|b_{i1}, b_{i2})) = \boldsymbol{\Sigma}_{i2}(v_i) + \psi_2^2(1-\rho^2)\, \mathbf{J}_{n_{i2}}. \end{aligned}$$

This establishes the result. □

	setting		
parameter	1	2	3
(β_1, β_2)	(185, 200)	(185, 190)	(185, 185)
$(\log\psi_1^2, \log\psi_2^2, \rho)$	(8, 9, 0.975)	(8, 8.25, 0.975)	(8, 8, 0.975)
$(\log\sigma_1^2, \log\sigma_2^2)$	*homoscedastic model,* $\delta = 0$		
	(1, 2)	(1, 1.25)	(1,1)
	heteroscedastic model, $\delta \in \{0.9, 1, 1.1\}$		
	(-9, -8)	(-9, -8.75)	(-9, -9)

Table 1: Parameter settings for the simulation study.

		bias				MSE			
set[1]	par[1]	LA-L	LA-Q	MA-M	MA-B	LA-L	LA-Q	MA-M	MA-B
1	β_1	-0.39	-0.29	-0.28	-0.28	51.74	49.72	55.20	55.20
	β_2	-0.54	-0.35	-0.35	-0.35	143.19	138.86	152.84	152.84
	$\log \sigma_1^2$	-1.10	-1.07	-1.11	-1.11	10.56	10.11	11.27	11.28
	$\log \sigma_2^2$	-0.31	-0.39	-0.35	-0.35	12.82	12.70	13.46	13.48
	$\log \psi_1^2$	-0.03	-0.03	-0.03	-0.03	0.04	0.04	0.04	0.04
	$\log \psi_2^2$	-0.03	-0.03	-0.03	-0.03	0.04	0.04	0.05	0.05
	$z(\rho)$	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.02
	δ_1	0.10	0.10	0.11	0.11	0.10	0.09	0.10	0.10
	δ_2	0.03	0.03	0.03	0.03	0.12	0.12	0.13	0.13
2	β_1	0.18	0.24	0.08	0.08	57.13	54.98	59.18	59.18
	β_2	0.08	0.14	-0.04	-0.04	73.32	70.12	76.01	76.01
	$\log \sigma_1^2$	-1.11	-1.14	-1.17	-1.17	11.98	12.24	12.79	12.81
	$\log \sigma_2^2$	-0.17	-0.17	-0.21	-0.21	12.55	12.32	13.26	13.28
	$\log \psi_1^2$	-0.05	-0.05	-0.05	-0.05	0.05	0.05	0.05	0.05
	$\log \psi_2^2$	-0.05	-0.05	-0.05	-0.05	0.04	0.04	0.04	0.04
	$z(\rho)$	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02
	δ_1	0.11	0.11	0.11	0.11	0.11	0.11	0.12	0.12
	δ_2	0.01	0.01	0.02	0.02	0.12	0.11	0.12	0.12
3	β_1	0.21	0.24	0.27	0.27	62.83	58.48	64.87	64.87
	β_2	0.25	0.25	0.29	0.29	62.76	58.70	64.24	64.24
	$\log \sigma_1^2$	-0.99	-0.92	-0.99	-0.99	12.22	11.88	12.58	12.60
	$\log \sigma_2^2$	-0.19	-0.19	-0.20	-0.20	11.24	10.92	11.43	11.45
	$\log \psi_1^2$	-0.04	-0.04	-0.04	-0.04	0.05	0.05	0.05	0.05
	$\log \psi_2^2$	-0.04	-0.04	-0.04	-0.04	0.05	0.05	0.05	0.05
	$z(\rho)$	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02
	δ_1	0.09	0.09	0.09	0.09	0.11	0.11	0.11	0.11
	δ_2	0.02	0.02	0.02	0.02	0.10	0.10	0.11	0.11

[1] set = setting, par=parameter.

Table 2: Estimated biases and MSE's of estimators computed using four model fitting methods when $\delta = 0$.

		bias				MSE			
set[1]	par[1]	LA-L	LA-Q	MA-M	MA-B	LA-L	LA-Q	MA-M	MA-B
1	β_1	0.16	0.10	-0.07	-0.07	54.02	54.17	62.19	62.20
	β_2	0.29	0.23	0.05	0.05	144.71	146.97	168.41	168.40
	$\log \sigma_1^2$	0.62	0.58	0.46	0.45	10.55	10.19	11.89	11.91
	$\log \sigma_2^2$	0.37	0.25	0.37	0.35	9.07	8.34	9.42	9.43
	$\log \psi_1^2$	-0.05	-0.05	-0.06	-0.06	0.04	0.04	0.04	0.04
	$\log \psi_2^2$	-0.05	-0.05	-0.05	-0.05	0.04	0.04	0.04	0.04
	$z(\rho)$	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02
	δ_1	-0.06	-0.06	-0.05	-0.05	0.10	0.10	0.11	0.11
	δ_2	-0.04	-0.03	-0.04	-0.04	0.08	0.08	0.09	0.09
2	β_1	0.39	0.40	0.27	0.28	51.36	46.01	56.70	56.70
	β_2	0.58	0.55	0.51	0.50	65.66	59.09	72.77	72.76
	$\log \sigma_1^2$	0.43	0.33	0.29	0.27	9.41	9.11	10.95	10.99
	$\log \sigma_2^2$	-0.02	-0.04	-0.03	-0.05	8.86	8.62	9.35	9.40
	$\log \psi_1^2$	-0.07	-0.06	-0.07	-0.07	0.04	0.04	0.04	0.04
	$\log \psi_2^2$	-0.07	-0.07	-0.07	-0.07	0.04	0.04	0.04	0.04
	$z(\rho)$	0.01	0.01	0.01	0.01	0.02	0.03	0.02	0.02
	δ_1	-0.04	-0.03	-0.03	-0.03	0.09	0.09	0.10	0.10
	δ_2	-0.003	-0.001	-0.002	0.000	0.08	0.08	0.09	0.09
3	β_1	0.65	0.53	0.34	0.34	49.96	49.58	54.30	54.31
	β_2	0.63	0.51	0.39	0.39	49.68	50.12	54.69	54.69
	$\log \sigma_1^2$	0.54	0.39	0.40	0.38	7.88	7.70	9.07	9.09
	$\log \sigma_2^2$	0.24	0.30	0.26	0.24	8.86	8.74	9.39	9.41
	$\log \psi_1^2$	-0.05	-0.05	-0.05	-0.05	0.04	0.04	0.04	0.04
	$\log \psi_2^2$	-0.05	-0.05	-0.05	-0.05	0.04	0.04	0.04	0.04
	$z(\rho)$	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02
	δ_1	-0.05	-0.04	-0.04	-0.04	0.08	0.07	0.08	0.08
	δ_2	-0.03	-0.03	-0.03	-0.03	0.08	0.08	0.09	0.09

[1] set = setting, par = parameter.

Table 3: Estimated biases and MSE's of estimators computed using four model fitting methods when $\delta = 1.1$.

		$\delta = 0$				$\delta = 1.1$			
set[1]	par[1]	LA-L	LA-Q	MA-M	MA-B	LA-L	LA-Q	MA-M	MA-B
1	β_1	94.2	95.2	95.2	95.2	91.4	92.0	94.0	94.0
	β_2	94.6	94.6	95.2	95.2	91.8	92.0	93.6	93.6
	$\log \sigma_1^2$	89.4	91.2	95.2	95.4	91.4	91.8	95.0	95.0
	$\log \sigma_2^2$	93.6	93.6	94.6	94.6	91.6	94.0	96.0	96.0
	$\log \psi_1^2$	92.8	92.8	93.6	93.6	90.2	90.0	93.0	93.0
	$\log \psi_2^2$	92.6	92.8	93.8	93.8	90.8	90.6	93.8	93.8
	$z(\rho)$	91.2	91.8	92.6	92.6	92.8	92.6	95.8	95.8
	δ_1	90.4	91.6	95.4	95.4	91.0	91.0	94.0	94.0
	δ_2	93.2	93.2	94.2	94.2	92.0	94.0	96.0	96.0
2	β_1	93.8	93.6	94.4	94.4	91.6	92.8	95.2	95.2
	β_2	93.4	93.8	94.0	94.0	92.6	92.6	95.2	95.2
	$\log \sigma_1^2$	88.0	88.6	94.6	94.4	90.4	91.2	96.6	96.2
	$\log \sigma_2^2$	93.2	92.6	93.8	93.8	92.0	93.4	97.6	97.6
	$\log \psi_1^2$	92.2	91.8	93.0	93.0	89.6	89.8	93.6	93.6
	$\log \psi_2^2$	92.4	92.2	93.4	93.4	89.6	89.4	93.8	93.8
	$z(\rho)$	94.6	94.4	95.8	95.8	90.0	90.0	94.0	93.8
	δ_1	87.4	88.6	94.6	94.6	90.4	91.2	96.0	96.0
	δ_2	93.6	93.0	94.2	94.2	91.8	93.8	97.8	97.8
3	β_1	92.6	93.6	93.2	93.2	91.2	91.4	93.6	93.6
	β_2	92.6	93.8	93.6	93.6	90.8	91.0	93.4	93.4
	$\log \sigma_1^2$	88.0	89.2	94.0	94.0	92.0	93.2	97.4	97.4
	$\log \sigma_2^2$	96.4	97.0	97.0	97.0	93.4	94.8	97.6	97.6
	$\log \psi_1^2$	92.2	92.6	93.0	93.0	91.2	90.8	94.4	94.4
	$\log \psi_2^2$	92.6	92.4	93.4	93.4	90.6	90.6	93.6	93.4
	$z(\rho)$	93.4	93.6	94.4	94.4	89.0	88.6	92.4	92.4
	δ_1	87.8	89.2	94.0	94.0	92.0	93.4	97.8	97.8
	δ_2	95.8	96.4	96.8	96.6	92.8	93.6	96.8	96.8

[1] set = setting, par = parameter.

Table 4: Estimated coverage probabilities (in %) of 95% confidence intervals computed using four model fitting methods.

	likelihood ratio test				score test			
setting	LA-L	LA-Q	MA-M	MA-B	LA-L	LA-Q	MA-M	MA-B
1	3.6	3.6	4.0	4.0	0.4	5.4	6.0	6.0
2	4.6	4.8	5.4	5.2	0.4	6.0	6.2	6.2
3	3.4	2.8	3.2	3.2	0.6	4.0	4.6	4.6

Table 5: Estimated type I error probabilities (in %) for 5% level likelihood ratio test and score test of homoscedasticity performed using four model fitting methods.

	estimation method	
	LA-L/LA-Q	MA-M/MA-B
parameter	estimate (SE)	estimate (SE)
β_1	184.38 (6.53)	184.38 (6.54)
β_2	189.98 (6.66)	189.98 (6.66)
$\log \sigma_1^2$	-9.34 (0.57)	-9.43 (0.57)
$\log \sigma_2^2$	-8.34 (0.58)	-8.57 (0.59)
$\log \psi_1^2$	8.36 (0.14)	8.36 (0.14)
$\log \psi_2^2$	8.40 (0.14)	8.40 (0.14)
$z(\rho)$	2.91 (0.10)	2.91 (0.10)
δ_1	1.00 (0.05)	1.02 (0.06)
δ_2	0.96 (0.06)	0.99 (0.06)

Table 6: Estimates and their standard errors (SE's) for cholesterol data computed using four estimation methods. The two likelihood approximation methods and the two model approximation methods produce identical estimates when rounded to two decimal places.

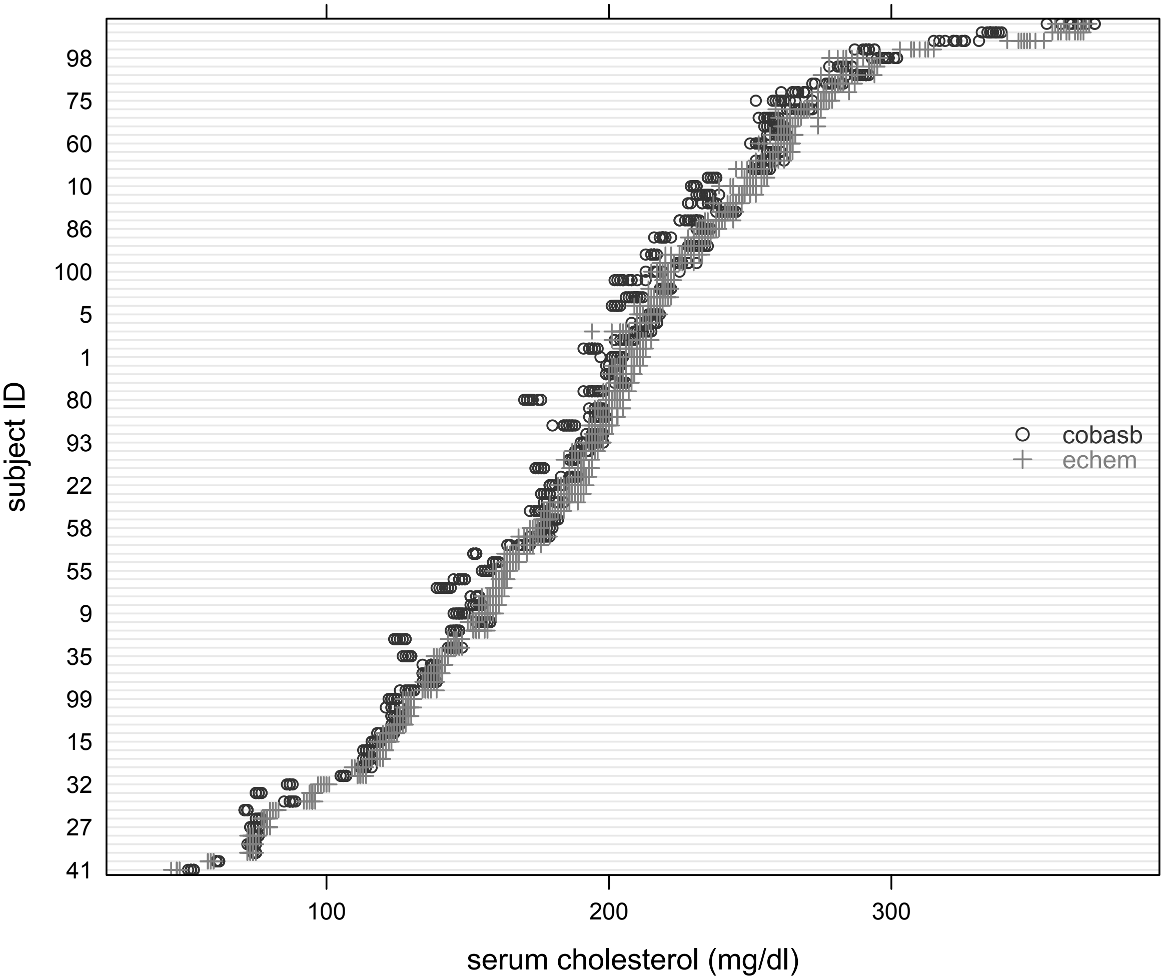

Figure 1: A trellis plot of cholesterol measurements using two assays — Cobas Bio (cobasb) and Ekatachem 700 (echem).

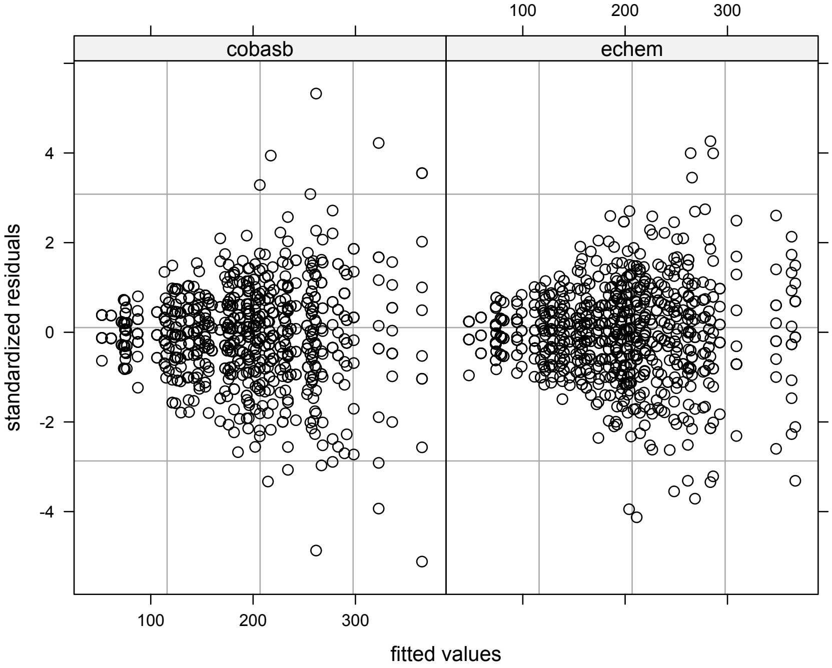

Figure 2: Separate residual plot for each assay when the homoscedastic model is fit to the cholesterol data.

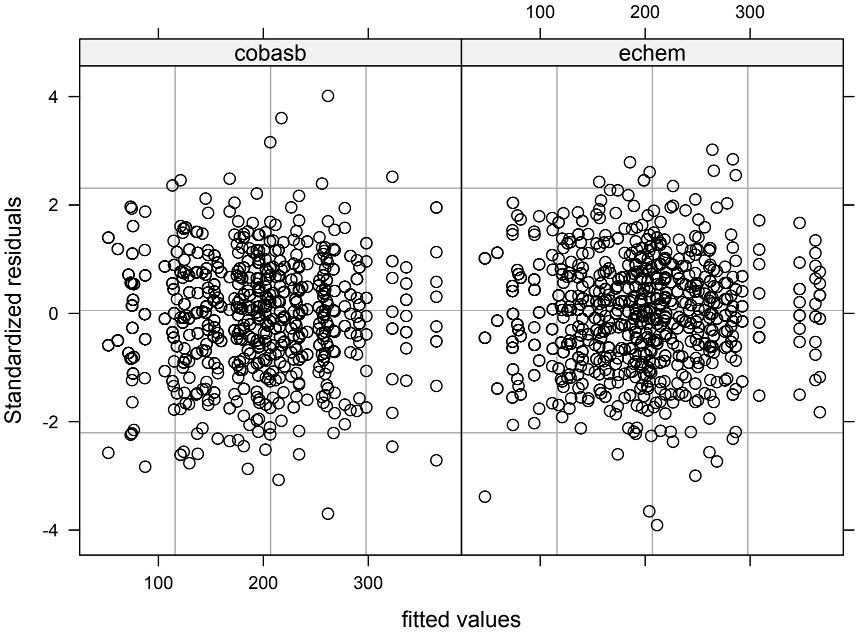

Figure 3: Separate residual plot for each assay when a heteroscedastic model is fit to the cholesterol data

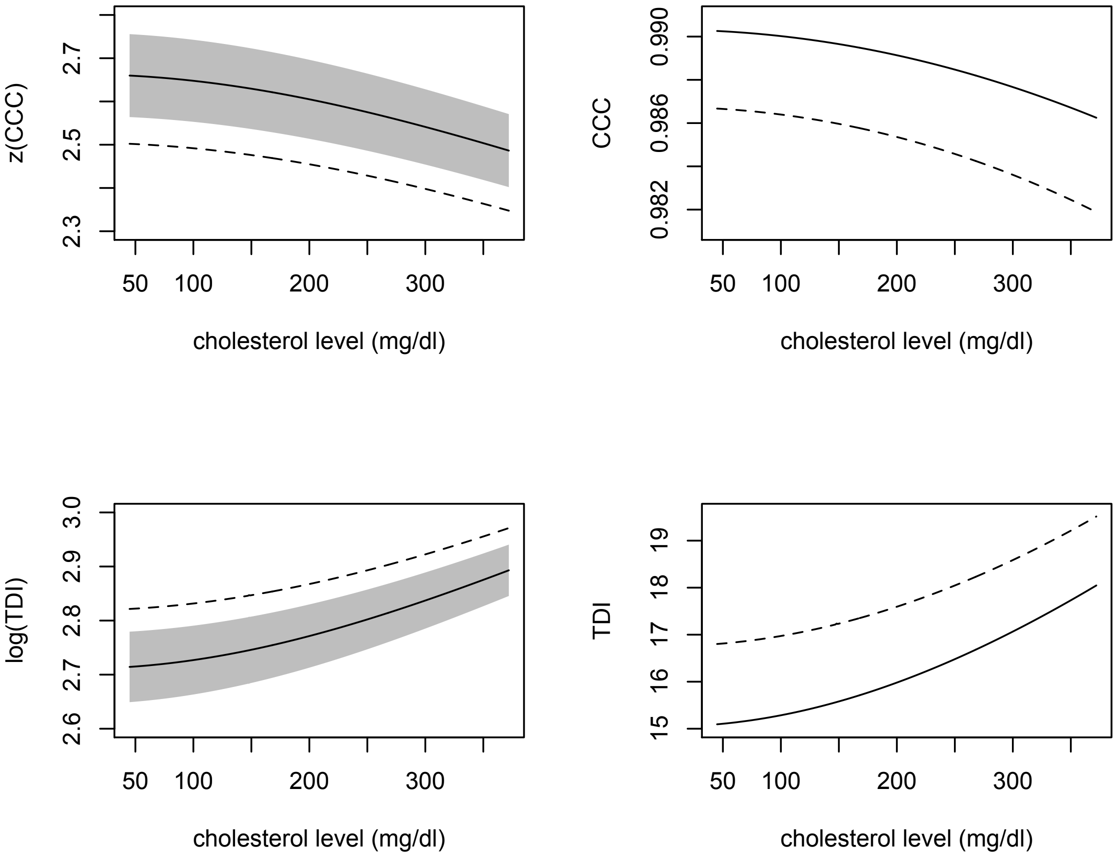

Figure 4: Estimates (solid line) and 95% pointwise one-sided confidence bands (broken line) for two agreement measures and their transformations for cholesterol data. A lower band is presented for CCC and its Fisher's z-transformation and an upper band is presented for TDI (with $p = 0.90$) and its log transformation. The shaded region in a plot represents estimate ± standard error.