A Heteroscedastic Measurement Error Model for Method Comparison Data With Replicate Measurements

Lakshika S. Nawarathna

Department of Statistics and Computer Science

University of Peradeniya

Peradeniya 20400, Sri Lanka

Pankaj K. Choudhary[1]

Department of Mathematical Sciences, FO 35

University of Texas at Dallas

Richardson, TX 75083-0688, USA

Abstract

Measurement error models offer a flexible framework for modeling data collected in studies comparing methods of quantitative measurement. These models generally make two simplifying assumptions: (a) the measurements are homoscedastic; and (b) the unobservable true values of the methods are linearly related. One or both of these assumptions may be violated in practice. In particular, error variabilities of the methods may depend on the magnitude of measurement or the true values may be nonlinearly related. Data with these features call for a heteroscedastic measurement error model that allows nonlinear relationships in the true values. We present such a model for the case when the measurements are replicated, discuss its fitting, and explain how to evaluate similarity of measurement methods and agreement between them, which are two common goals of data analysis, under this model. Model fitting involves dealing with lack of a closed form for the likelihood function. We consider estimation

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

methods that approximate either the likelihood or the model to yield approximate maximum likelihood estimates. The fitting methods are evaluated in a simulation study. The proposed methodology is used to analyze a cholesterol dataset.

Keywords: Agreement, calibration, mixed-effects model, nonlinear model, repeated measures, total deviation index.

1 Introduction

Method comparison studies are concerned with comparing a new cheaper or easier test method for measuring a quantitative variable with an established reference method. Such studies are routinely conducted in biomedical disciplines. The variable being measured often has some clinical interest, e.g., cholesterol level or fat content. The methods may be medical devices, assays, clinical observers, or measurement protocols. None of the methods is considered error-free. The data consist of at least one measurement from each method on every subject in the study. Our focus is on the study design wherein measurements from both methods are replicated. The primary goal of the comparison, especially if the methods measure in the same nominal unit, is to evaluate agreement between the methods to see if they can be used interchangeably. A number of articles have developed statistical methodologies for this purpose, including [1–6]. See [7, 8] for overviews. Another common goal of the comparison, irrespective of whether the methods measure in the same nominal unit or not, is to evaluate similarity of the methods by comparing their accuracies and precisions, and recalibrate one method with respect to the other. Statistical methodologies for accomplishing this goal are reviewed in [9].

Regardless of the goal of a method comparison study, modeling of data is a key step in the data analysis. Two modeling frameworks, namely, a mixed-effects model [10] and a mea-

surement error model [11], are especially popular. A mixed-effects model is employed when the methods can be assumed to have the same measurement scale, meaning that the true (i.e., error-free) values of the methods may differ only by a constant [12–16]. An example of methods with the same scales is two thermometers, one measuring in Celsius (°C) and the other in Kelvin (K) because $\mathrm{K} - {}^{\circ}\mathrm{C} = 273.15$. The assumption of a common scale is not needed in a measurement error model because it allows the true values to be linearly related rather than just differ by a constant [17–20]. An example of methods with linearly related true values is two thermometers, one measuring in °C and the other in Fahrenheit (°F) because ${}^{\circ}\mathrm{F} = 32 + (9/5)\,{}^{\circ}\mathrm{C}$. Note that for methods to have the same scale it is neither necessary nor sufficient that they have the the same unit of measurement. While thermometers measuring in °C and K are an example of methods with different units but same scales, an example of methods with same units but different scales is two thermometers, both measuring in °C, but one consistently giving 10% higher measurement than the other due to miscalibration. Of course, in these temperature related examples we *know* the relationships between the various thermometers. But in most method comparison studies in practice these relationships need to be estimated from the data.

Obviously since a constant difference in true values is a special case of a linear relation in them when the slope is one, a measurement error model offers a more flexible framework for modeling method comparison data than a mixed-effects model. Measurement error models have been advocated in [17–21]. But these models generally make a simplifying assumption that the measurements are homoscedastic, i.e., the variability of measurements remains constant over the entire measurement range. In practice, however, it frequently happens that the variability of a measurement changes with its magnitude [1, 17, 22]. The cholesterol data of [12], which motivated this work and is analyzed later in this article, provides a specific example of this phenomenon. In presence of such heteroscedasticity, the precisions of the methods

as well as the extent of agreement between them change with the magnitude of measurement. But these quantities would be treated as constants if the heteroscedasticity is ignored, leading to potentially misleading conclusions. Variance stabilizing transformation of data may be used to remove the heteroscedasticity, but the difference of transformed measurements may be difficult to interpret. This is a problem because the measurement differences need to be interpretable to evaluate agreement between the methods [22]. Therefore, models that explicitly incorporate heteroscedasticity are of considerable interest to practitioners.

Recently, a heteroscedastic mixed-effects model was proposed in [23] for replicated method comparison data. As for the measurement error model framework, heteroscedastic models have been considered, e.g., in [9, 24, 25], but none is specifically designed for replicated method comparison data. This brings us to the main goals of this article, which are to present such a model, discuss computational algorithms for fitting it, and illustrate its application. The novelty of our approach also lies in that we allow the true values of the methods to be nonlinearly related thereby obtaining the standard model with linear relationship as an important special case. Heteroscedastic models allowing nonlinear relationships have hitherto not been studied in the method comparison literature.

The rest of this article is organized as follows. Section 2 presents the proposed model. Section 3 discusses computational methods for fitting it. Section 4 describes a simulation study to evaluate the model fitting methods. Section 5 shows how to use the model to evaluate similarity of measurement methods and agreement between them. Section 6 illustrates an application by analyzing a cholesterol data set. Section 7 concludes with a discussion. All the computations in this article have been performed using the statistical software `R` [26].

2 The proposed heteroscedastic model

Consider a method comparison study involving two measurement methods and m subjects. Let Y_{ijk} be the kth replicate measurement by the jth method on the ith subject. The data in the study consist of Y_{ijk}, $k = 1, \ldots, n_{ij}$, $j = 1, 2$, $i = 1, \ldots, m$. Here method 1 represents the reference method and method 2 represents the test method. The multiple measurements from a method on a subject are exchangeable in that they are replications of the same underlying measurement. The replicate measurements from the two methods on a subject are dependent but they are not paired. In fact, the methods may not even have the same number of replications on a subject. Let $n_i = n_{i1} + n_{i2}$ be the total number of measurements on the ith subject. The n_i need not be equal. In what follows, we will use bold-face letters to denote vectors and matrices. By default, a vector is a column vector unless specified otherwise. The transpose of a vector or matrix $\mathbf{A}$ is denoted as $\mathbf{A}^T$.

Let $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijn_{ij}})^T$ denote the n_{ij}-vector of measurements on subject i from method j. The n_i-vector $\mathbf{Y}_i = (\mathbf{Y}_{i1}^T, \mathbf{Y}_{i2}^T)^T$ denotes all measurements on subject i. Let $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \tilde{Y}_2)^T$ denote paired measurements from the two methods on a randomly selected subject from the population. We think of $\tilde{\mathbf{Y}}$ as a "typical" measurement pair in that the observed (Y_{i1k}, Y_{i2l}) pairs — even though the replications are not paired by design — are identically distributed as $\tilde{\mathbf{Y}}$.

Let $\boldsymbol{\theta}$ be the vector of all unknown model parameters. We use $h_{\boldsymbol{\theta}}(\mathbf{y}_1, \mathbf{y}_2)$ for the joint probability density function of $(\mathbf{Y}_1, \mathbf{Y}_2)$ and $h_{\boldsymbol{\theta}}(\mathbf{y}_1|\mathbf{y}_2)$ for the conditional density of $\mathbf{Y}_1|\mathbf{Y}_2 = \mathbf{y}_2$. We also use $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote a d-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

2.1 The model for $\tilde{\mathbf{Y}}$

To prepare the groundwork for presenting a heteroscedastic model for the observed data, we first present it for $\tilde{\mathbf{Y}}$ and then adapt it for the observed data. Let b denote the true unobservable measurement underlying $\tilde{\mathbf{Y}}$, and e_1 and e_2 denote the random errors of the two methods. The basic measurement error model for $\tilde{\mathbf{Y}}$ is written as [18]

$$\tilde{Y}_1 = b + e_1,\ \tilde{Y}_2 = \beta_0 + \beta_1 b + e_2;\quad b \sim \mathcal{N}_1(\mu, \tau^2),\ e_1 \sim \mathcal{N}_1(0, \sigma_1^2),\ e_2 \sim \mathcal{N}_1(0, \sigma_2^2), \qquad (1)$$

where β_0 and β_1 are regression coefficients respectively known as *fixed* and *proportional* biases of the test method, and (b, e_1, e_2) are mutually independent. For reasons of model identifiability, the true measurement b is also the true value of the reference method. This model postulates a linear relationship between the true values b and $\beta_0 + \beta_1 b$ of the two methods. The methods have different scales when the slope $\beta_1 \neq 1$. The model (1) is called a "measurement error" model because the covariate b of $\tilde{Y}_2$ is measured with error as $\tilde{Y}_1$.

We now make three changes to this basic model. First, we replace the linear calibration function $\beta_0 + \beta_1 b$ relating the true values of the two methods by a more general function $f(b, \boldsymbol{\beta})$. The function f has a known parametric form which depends on a fixed unknown parameter vector $\boldsymbol{\beta}$. Moreover, f is differentiable and may be nonlinear in b as well as in $\boldsymbol{\beta}$. Specific examples of f include $\beta_0 + \beta_1 b$ (linear model), $\beta_0 + \beta_1 b + \beta_2 b^2$ (quadratic model), and $\beta_0 \exp(-\beta_1 b)$ (exponential model).

Second, we add independent method $\times$ subject interaction effects — $b_1 \sim \mathcal{N}_1(0, \psi^2)$ and $b_2 \sim \mathcal{N}_1(0, \psi^2)$ — to the respective expressions for $\tilde{Y}_1$ and $\tilde{Y}_2$. These effects are essentially subject specific biases of the methods. They are also known as "equation errors" in the measurement error literature and as "matrix effects" in analytical chemistry [9]. They appear additively in the model and are mutually independent of (b, e_1, e_2). See also the discussion in Section 7 for a note regarding how the equal variance assumption for the two interaction

effects may be relaxed. It may be noted that interaction effects for both methods are almost always included in the model when a mixed-effects model is used for replicated method comparison data. However, when a measurement error model is used, they are often included only for the test method but not for the reference method, see, e.g., [9, pp. 74-77].

Finally, we replace the constant error variance σ_j^2 by $\text{var}[e_j|b] = \sigma_j^2\, g_j^2(b, \boldsymbol{\delta}_j),\ j = 1, 2$, where g_j is a *variance function* that models how the error variance of the jth method depends on the true value b. This g_j is also differentiable and has a known parametric form depending on an unknown heteroscedasticity parameter vector $\boldsymbol{\delta}_j$, which is such that $g_j(b, \boldsymbol{\delta}_j) \equiv 1$ when $\boldsymbol{\delta}_j = 0$. The two methods may have different variance functions. Examples of a variance function include $|b|^\delta$ (power model), $\delta_0 + |b|^{\delta_1}$ (constant plus power model), and $\exp(\delta b)$ (exponential model) [10]. The model becomes homoscedastic when $\boldsymbol{\delta}_1 = \mathbf{0} = \boldsymbol{\delta}_2$.

After these changes, the basic model (1) for $\tilde{\mathbf{Y}}$ becomes a heteroscedastic measurement error model,

$$\begin{aligned} &\tilde{Y}_1 = b + b_1 + e_1,\ \tilde{Y}_2 = f(b, \boldsymbol{\beta}) + b_2 + e_2; \\ &b_j \sim \mathcal{N}_1(0, \psi^2),\ e_j|b \sim \mathcal{N}_1\big(0, \sigma_j^2\, g_j^2(b, \boldsymbol{\delta}_j)\big),\ b \sim \mathcal{N}_1(\mu, \tau^2),\ j = 1, 2. \end{aligned} \tag{2}$$

Here e_1 and e_2 are conditionally independent given b. Marginally, they are uncorrelated but dependent. Further, b is independent of (b_1, b_2) and (b_1, b_2, e_1, e_2) are mutually independent. This model is nonlinear in b unless f is linear in b and the g_j are free of b. By marginalizing over (b_1, b_2), we get a hierarchical representation of this model as

$$\tilde{Y}_1|b \sim \mathcal{N}_1\left(b, \psi^2 + \sigma_1^2\, g_1^2(b, \boldsymbol{\delta}_1)\right),\ \tilde{Y}_2|b \sim \mathcal{N}_1\big(f(b, \boldsymbol{\beta}), \psi^2 + \sigma_2^2\, g_2^2(b, \boldsymbol{\delta}_2)\big),\ b \sim \mathcal{N}_1(\mu, \tau^2), \tag{3}$$

where $\tilde{Y}_1$ and $\tilde{Y}_2$ are conditionally independent given b. In general, further marginalization over b does not yield a closed-form marginal distribution for $\tilde{\mathbf{Y}}$, albeit its marginal mean

vector and covariance matrix can be written as

$$E[\tilde{\mathbf{Y}}] = \begin{bmatrix} \mu \\ E[f(b,\boldsymbol{\beta})] \end{bmatrix},$$

$$\text{var}[\tilde{\mathbf{Y}}] = \text{diag}\big\{\psi^2 + \sigma_1^2\, E[g_1^2(b,\boldsymbol{\delta}_1)], \psi^2 + \sigma_2^2\, E[g_2^2(b,\boldsymbol{\delta}_2)]\big\} + \boldsymbol{\Gamma}, \tag{4}$$

where $\boldsymbol{\Gamma}$ is the covariance matrix of $(b, f(b,\boldsymbol{\beta}))^T$,

$$\boldsymbol{\Gamma} = \begin{bmatrix} \tau^2 & \text{cov}[b, f(b,\boldsymbol{\beta})] \\ \text{cov}[b, f(b,\boldsymbol{\beta})] & \text{var}[f(b,\boldsymbol{\beta})] \end{bmatrix}. \tag{5}$$

2.2 The model for observed data

We get a model for observed data $\mathbf{Y}_i$ from that of $\tilde{\mathbf{Y}}$ in (2) by simply replacing $\tilde{Y}_j$ with Y_{ijk} and (b, b_j, e_j) with its independent copies (b_i, b_{ij}, e_{ijk}) for $k = 1, \ldots, n_{ij}$, $j = 1, 2$, $i = 1, \ldots, m$. This gives

$$Y_{i1k} = b_i + b_{i1} + e_{i1k},\ Y_{i2k} = f(b_i,\boldsymbol{\beta}) + b_{i2} + e_{i2k};$$

$$b_{ij} \sim \mathcal{N}_1(0,\psi^2),\ e_{ijk}|b_i \sim \mathcal{N}_1\big(0, \sigma_j^2\, g_j^2(b_i,\boldsymbol{\delta}_j)\big),\ b_i \sim \mathcal{N}_1(\mu,\tau^2). \tag{6}$$

In this model, the multiple measurements from method j on subject i are dependent because they share the same b_i and b_{ij}. Furthermore, the measurements from different methods on subject i are also dependent because they share the same b_i. The measurements on different subjects are independent.

To write this model in the matrix form, let $\mathbf{1}_n$ and $\mathbf{0}_n$ be n-vectors of ones and zeros, and define $\mathbf{e}_{ij} = (e_{ij1}, \ldots, e_{ijn_{ij}})^T$,

$$\mathbf{e}_i = \begin{bmatrix} \mathbf{e}_{i1} \\ \mathbf{e}_{i2} \end{bmatrix}, \mathbf{b}_i = \begin{bmatrix} b_{i1} \\ b_{i2} \end{bmatrix}, \mathbf{U}_i = \begin{bmatrix} \mathbf{1}_{n_{i1}} \\ \mathbf{0}_{n_{i2}} \end{bmatrix}, \mathbf{V}_i = \begin{bmatrix} \mathbf{0}_{n_{i1}} \\ \mathbf{1}_{n_{i2}} \end{bmatrix}, \mathbf{Z}_i = [\mathbf{U}_i, \mathbf{V}_i]. \tag{7}$$

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

$$\boldsymbol{\Sigma}_{ij}(b) = \text{diag}\big\{\sigma_j^2\, g_j^2(b,\boldsymbol{\delta}_j), \ldots, \sigma_j^2\, g_j^2(b,\boldsymbol{\delta}_j)\big\},\ \boldsymbol{\Sigma}_i(b) = \text{diag}\big\{\boldsymbol{\Sigma}_{i1}(b), \boldsymbol{\Sigma}_{i2}(b)\big\}.$$

Now the model (6) can be written in the matrix form as

$$\mathbf{Y}_i = \mathbf{U}_i b_i + \mathbf{V}_i f(b_i, \boldsymbol{\beta}) + \mathbf{Z}_i \mathbf{b}_i + \mathbf{e}_i;$$

$$\mathbf{e}_i | b_i \sim \mathcal{N}_{n_i}\big(\mathbf{0}, \boldsymbol{\Sigma}_i(b_i)\big),\ \mathbf{b}_i \sim \mathcal{N}_2(\mathbf{0}, \psi^2 \text{diag}\{1,1\}),\ b_i \sim \mathcal{N}_1(\mu, \tau^2),\ i = 1, ..., m. \tag{8}$$

Proceeding as in (3), we can represent the model in a hierarchical manner as

$$\mathbf{Y}_i | b_i \sim \mathcal{N}_{n_i}\big(\mathbf{U}_i b_i + \mathbf{V}_i f(b_i, \boldsymbol{\beta}), \psi^2 \mathbf{Z}_i \mathbf{Z}_i^T + \boldsymbol{\Sigma}_i(b_i)\big),\ b_i \sim \mathcal{N}_1(\mu,\, \tau^2). \tag{9}$$

Here $\mathbf{Y}_{i1}$ and $\mathbf{Y}_{i2}$ are conditionally independent given b_i because the matrix

$$\psi^2 \mathbf{Z}_i \mathbf{Z}_i^T + \boldsymbol{\Sigma}_i(b_i) = \text{diag}\big\{\psi^2 \mathbf{1}_{n_{i1}} \mathbf{1}_{n_{i1}}^T + \boldsymbol{\Sigma}_{i1}(b_i),\ \psi^2 \mathbf{1}_{n_{i2}} \mathbf{1}_{n_{i2}}^T + \boldsymbol{\Sigma}_{i2}(b_i)\big\}$$

has a diagonal structure. This is expected since the dependence in $\mathbf{Y}_{i1}$ and $\mathbf{Y}_{i2}$ is induced only through the common b_i. Moreover, analogous to (4), the marginal mean vector and covariance matrix of $\mathbf{Y}_i$ are

$$E[\mathbf{Y}_i] = \mathbf{U}_i \mu + \mathbf{V}_i E[f(b, \boldsymbol{\beta})],\quad \text{var}[\mathbf{Y}_i] = \mathbf{Z}_i \boldsymbol{\Gamma} \mathbf{Z}_i^T + \psi^2 \mathbf{Z}_i \mathbf{Z}_i^T + E[\boldsymbol{\Sigma}_i(b)], \tag{10}$$

with $\boldsymbol{\Gamma}$ given by (5). In principle, the marginal probability density function $h_{\boldsymbol{\theta}}(\mathbf{y}_i)$ of $\mathbf{Y}_i$ can be obtained as

$$h_{\boldsymbol{\theta}}(\mathbf{y}_i) = \int_{-\infty}^{\infty} h_{\boldsymbol{\theta}}(\mathbf{y}_i, b_i) db_i. \tag{11}$$

Here $h_{\boldsymbol{\theta}}(\mathbf{y}_i, b_i) = h_{\boldsymbol{\theta}}(\mathbf{y}_{i1}|b_i) h_{\boldsymbol{\theta}}(\mathbf{y}_{i2}|b_i) h_{\boldsymbol{\theta}}(b_i)$ from conditional independence. The densities involved in this expression are normal densities obtained from (9). However, the integral (11) does not have a closed-form in general.

2.3 The case of linear f

The case of linear calibration function $f(b, \boldsymbol{\beta}) = \beta_0 + \beta_1 b$ is of special interest in practice. In this case, the moments of $\tilde{\mathbf{Y}}$ in (4) and $\mathbf{Y}_i$ in (10) simplify because we explicitly have

$$E[f(b, \boldsymbol{\beta})] = \beta_0 + \beta_1 \mu,\quad \boldsymbol{\Gamma} = \begin{bmatrix} \tau^2 & \beta_1 \tau^2 \\ \beta_1 \tau^2 & \beta_1^2 \tau^2 \end{bmatrix}.$$

In addition, if the model is homoscedastic then from (9), $\mathbf{Y}_i \sim \mathcal{N}_{n_i}(E[\mathbf{Y}_i], \text{var}[\mathbf{Y}_i])$, where

$$\begin{aligned} E[\mathbf{Y}_i] &= \mathbf{V}_i\beta_0 + (\mathbf{U}_i + \mathbf{V}_i\beta_1)\mu, \\ \text{var}[\mathbf{Y}_i] &= \tau^2(\mathbf{U}_i + \mathbf{V}_i\beta_1)(\mathbf{U}_i + \mathbf{V}_i\beta_1)^T + \psi^2\mathbf{Z}_i\mathbf{Z}_i^T + \text{diag}\{\sigma_1^2\mathbf{1}_{n_{i1}}^T, \sigma_2^2\mathbf{1}_{n_{i2}}^T\}. \end{aligned} \tag{12}$$

Thus, in this case the marginal density $h_{\boldsymbol{\theta}}(\mathbf{y}_i)$ is a normal density. This exception occurs because b_i appears in the model linearly, allowing it to be explicitly integrated out in (11).

3 Model fitting by maximum likelihood

3.1 Likelihood computation

Let $L(\boldsymbol{\theta})$ denote the likelihood function of parameter vector $\boldsymbol{\theta} = (\mu, \tau^2, \boldsymbol{\beta}^T, \psi^2, \sigma_1^2, \sigma_2^2, \boldsymbol{\delta}_1^T, \boldsymbol{\delta}_2^T)^T$ under model (6). By definition,

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

but $h_{\boldsymbol{\theta}}(\mathbf{y}_i)$, given by (11), does not have an explicit expression in general. Therefore, we now describe two approaches for computing it. The first numerically approximates the integral (11), whereas the second approximates the original model (6) so that the resulting density has a closed-form.

3.1.1 Approach 1: Numerical integration

Let $l_{\boldsymbol{\theta}}(\mathbf{y}_i, b_i) = -\log h_{\boldsymbol{\theta}}(\mathbf{y}_i, b_i)$ be the negative logarithm of the integrand in (11); $b_{i,\min}$ be its minimizer with respect to b_i; and $l''_{\boldsymbol{\theta}}(\mathbf{y}_i, b_{i,\min}) = (\partial^2/\partial b_i^2) l_{\boldsymbol{\theta}}(\mathbf{y}_i, b_i)|_{b_i = b_{i,\min}}$ be the corresponding Hessian at the minima. A simple approximation of the integral (11) is the *Laplace approximation* (LA) [27],

$$h_{\boldsymbol{\theta}}(\mathbf{y}_i) \approx (2\pi)^{1/2} \, |l''_{\boldsymbol{\theta}}(\mathbf{y}_i, b_{i,\min})|^{-1/2} \, h_{\boldsymbol{\theta}}(\mathbf{y}_i, b_{i,\min}).$$

Another approximation is given by the *Gauss-Hermite quadrature* (GH) [27]. To describe this, let $z_1, \ldots, z_M$ be the nodes and $w_1, \ldots, w_M$ be the associated quadrature weights with kernel $\exp(-z^2)$. The nodes are centered and scaled to achieve greater accuracy as [28]

$$c_{ir} = b_{i,\min} + 2^{1/2} l''_{\boldsymbol{\theta}}(\mathbf{y}_i, b_{i,\min})^{-1/2} z_r, \quad r = 1, \ldots, M.$$

The approximated integral in this case is

$$h_{\boldsymbol{\theta}}(\mathbf{y}_i) \approx 2^{1/2} \left| l''_{\boldsymbol{\theta}}(\mathbf{y}_i, b_{i,\min}) \right|^{-1/2} \sum_{r=1}^{M} h_{\boldsymbol{\theta}}(\mathbf{y}_i, \mathbf{c}_{ir}) w_r \exp(z_r^2).$$

This method reduces to LA when $M = 1$ because the sole node in this case is zero with weight $\pi^{1/2}$ [28]. This makes it clear that GH is not only more accurate but also more computationally demanding than LA. In practice, 20-30 nodes tend to provide reasonably good accuracy for GH method.

3.1.2 Approach 2: Model approximation by linearization

This approach approximates the model (6) by linearizing f and g_j functions in b_i — an unobservable random quantity — around an observable non-random quantity b_i^* that is close to b_i but is held fixed in model fitting. This kind of linearization is a standard strategy in fitting of nonlinear mixed-effects and measurement error models [10, 29] and even generalized linear mixed-effects models [30]. Expanding f and g_j functions around $b_i = b_i^*$ using Taylor series and keeping the first two terms for f and only the first term for g_j, we get

$$f(b_i, \boldsymbol{\beta}) \approx f(b_i^*, \boldsymbol{\beta}) + (b_i - b_i^*) f'(b_i^*, \boldsymbol{\beta}), \quad g_j(b_i, \boldsymbol{\delta}_j) \approx g_j(b_i^*, \boldsymbol{\delta}_j), \ j = 1, 2, \tag{13}$$

where $f'(b_i^*, \boldsymbol{\beta}) = (\partial / \partial b_i) f(b_i, \boldsymbol{\beta})|_{b_i = b_i^*}$. The approximation for f is exact when f is linear in b_i, whereas the approximation for g_j is exact when the model is homoscedastic. Replacing

f and g_j in (6) by their approximations in (13) gives the linearized version of (6) as

$$
\begin{aligned}
&Y_{i1k} = b_i + b_{i1} + e_{i1k},\ Y_{i2k} \approx f(b_i^*, \boldsymbol{\beta}) + (b_i - b_i^*) f'(b_i^*, \boldsymbol{\beta}) + b_{i2} + e_{i2k};\\
&b_{ij} \sim \mathcal{N}_1(0, \psi^2),\ e_{ijk} \dot{\sim} \mathcal{N}_1\big(0, \sigma_j^2\, g_j^2(b_i^*, \boldsymbol{\delta}_j)\big),\\
&b_i \sim \mathcal{N}_1(\mu, \tau^2),\quad k = 1, \ldots, n_{ij},\ j = 1, 2,\ i = 1, \ldots, m.
\end{aligned} \tag{14}
$$

Here the notation "$\dot{\sim}$" means "is approximately distributed as," and the approximation is caused by the linearization. Letting

$$
d(b_i^*, \boldsymbol{\beta}) = f(b_i^*, \boldsymbol{\beta}) - b_i^* f'(b_i^*, \boldsymbol{\beta}), \tag{15}
$$

the approximate model (14) can be written in the matrix form as

$$
\begin{aligned}
&\mathbf{Y}_i \approx \mathbf{V}_i d(b_i^*, \boldsymbol{\beta}) + (\mathbf{U}_i + \mathbf{V}_i f'(b_i^*, \boldsymbol{\beta}))b_i + \mathbf{Z}_i \mathbf{b}_i + \mathbf{e}_i;\\
&\mathbf{e}_i \dot{\sim} \mathcal{N}_{n_i}\big(\mathbf{0}, \boldsymbol{\Sigma}_i(b_i^*)\big),\ \mathbf{b}_i \sim \mathcal{N}_2(\mathbf{0}, \psi^2 \mathrm{diag}\{1, 1\}),\ b_i \sim \mathcal{N}_1(\mu, \tau^2),\ i = 1, ..., m.
\end{aligned} \tag{16}
$$

It follows from marginalizing over b_i and $\mathbf{b}_i$ that $\mathbf{Y}_i \dot{\sim} \mathcal{N}_{n_i}(E[\mathbf{Y}_i], \mathrm{var}[\mathbf{Y}_i])$, where

$$
\begin{aligned}
E[\mathbf{Y}_i] &\approx \mathbf{V}_i d(b_i^*, \boldsymbol{\beta}) + (\mathbf{U}_i + \mathbf{V}_i f'(b_i^*, \boldsymbol{\beta}))\mu,\\
\mathrm{var}[\mathbf{Y}_i] &\approx \tau^2 (\mathbf{U}_i + \mathbf{V}_i f'(b_i^*, \boldsymbol{\beta}))(\mathbf{U}_i + \mathbf{V}_i f'(b_i^*, \boldsymbol{\beta}))^T + \psi^2 \mathbf{Z}_i \mathbf{Z}_i^T + \boldsymbol{\Sigma}(b_i^*).
\end{aligned} \tag{17}
$$

Thus, $h_{\boldsymbol{\theta}}(\mathbf{y}_i)$ can be approximated by the density of this normal distribution. We refer to this model approximation method as MA. This closed-form approximation is made possible by the way f and g_j functions are linearized, which ensures that b_i appears in the model linearly. This also explains why only the first term in the Taylor expansion was kept for g_j. One may think of this approximation method as a *pseudo-likelihood approach* because the true model is approximated by model that leads to a normal marginal likelihood.

To implement this method it remains to choose b_i^*. A natural choice is $b_i^* = \overline{\mathbf{y}}_{i1}$, the mean for the reference method. The resulting model (14) with b_i^* held fixed can be fit via maximum likelihood (ML). An alternative choice for b_i^* is the best linear unbiased predictor

of b_i. The model in this case needs to be fit by an iterative scheme because the predictor itself depends on unknown model parameters [10, 29]. Empirical results in [23] show that this additional complexity in model fitting is not worthwhile at least for method comparison studies because the differences in parameter estimates are negligible. Therefore, we only work with $b_i^* = \overline{\mathbf{y}}_{i1}$ in this article.

3.2 Inference on model parameters

The likelihood function approximated using either of the three methods — LA, GH and MA — can be maximized by an optimization routine, e.g., `optim` function in `R`, to compute approximate ML estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. Subsequent inference on $\boldsymbol{\theta}$ ignores the approximation in $\hat{\boldsymbol{\theta}}$ and employs the standard large-sample theory of ML estimators [31]. In particular, when m is large, the standard errors (SEs) of estimates and confidence intervals for parameters are obtained by approximating the distribution of $\hat{\boldsymbol{\theta}}$ by a normal distribution with mean $\boldsymbol{\theta}$ and the inverse of the observed information matrix $\mathbf{I} = -(\partial/\partial\boldsymbol{\theta}^2) \log L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ as the covariance matrix. Here $L(\boldsymbol{\theta})$ represents the likelihood function under the model actually fit to the data. Moreover, the null hypothesis of homoscedasticity ($\boldsymbol{\delta}_1 = \mathbf{0} = \boldsymbol{\delta}_2$) is tested by performing a likelihood ratio test wherein the null distribution of the test statistic is approximated by a chi-square distribution with degrees of freedom equal to number of parameters set to zero under the null hypothesis. This strategy of ignoring the approximation in $\hat{\boldsymbol{\theta}}$ for further statistical inference is common in nonlinear mixed-effects and measurement error models [10, 29] and generalized linear mixed-effects models [30].

3.3 Fitted values and residuals

Let $\hat{\mathbf{Y}}_i$ denote the fitted value of $\mathbf{Y}_i$, $i = 1, \ldots, m$. Under the linearized model (16),

$$\hat{\mathbf{Y}}_i \approx \mathbf{V}_i d(b_i^*, \hat{\boldsymbol{\beta}}) + (\mathbf{U}_i + \mathbf{V}_i f'(b_i^*, \hat{\boldsymbol{\beta}}))\hat{b}_i + \mathbf{Z}_i \hat{\mathbf{b}}_i,$$

where $(\hat{b}_i, \hat{\mathbf{b}}_i)$ is the estimated best predictor of $(b_i, \mathbf{b}_i)$ given $\mathbf{Y}_i$, obtained by substituting $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ in

$$E[(b_i, \mathbf{b}_i)^T | \mathbf{Y}_i] \approx (\mu, 0, 0)^T + \text{diag}\{\tau^2, \psi^2, \psi^2\}(\text{var}[\mathbf{Y}_i])^{-1}(\mathbf{Y}_i - E[\mathbf{Y}_i]),$$

with $E[\mathbf{Y}_i]$ and $\text{var}[\mathbf{Y}_i]$ given by (17). The residuals can be computed as $\hat{\mathbf{e}}_i = \mathbf{Y}_i - \hat{\mathbf{Y}}_i$, $i = 1, \ldots, m$. These residuals and their standardized counterparts, computed by dividing residuals by estimated error standard deviations (SDs), are used for model checking.

3.4 Specifying f and g_j functions

Specifying the calibration function f and the variance functions g_j is a part of model building exercise which is no different from what we ordinarily do in regression modeling. Therefore, we proceed just the way we proceed to build a parametric regression model. In particular, this involves relying on graphical techniques, such as scatterplot of measurements from the two methods and the residual plot, to come up with preliminary forms for these functions. As we are dealing with unpaired replicate measurements here, we can plot either the randomly formed measurement pairs (Y_{i1k}, Y_{i2l}) [14] or the paired averages $(\overline{y}_{i1}, \overline{y}_{i2})$ on the scatterplot.

4 A simulation study

Our next task is to use Monte Carlo simulation to evaluate finite sample performance of the three model fitting methods — LA, GH, and MA — on four performance measures:

biases of parameter estimators, their mean squared errors (MSEs), coverage probabilities of 95% confidence intervals, and type I error probability for 5% level likelihood ratio test of homoscedasticity. We focus on $f(b, \boldsymbol{\beta}) = \beta_0 + \beta_1 b$ and $g_j(b, \boldsymbol{\delta}_j) = |b|^{\delta_j}$, $j = 1, 2$ because these are the functions we adopt later for analysis of cholesterol data. In addition, we assume a balanced design with $n_{ij} \in \{2, 3\}$ replications per method; let $\delta_1 = \delta_2 = \delta \in \{0, 0.9, 1, 1.1\}$; and take $m = 50$. Table 1 summarizes the actual parameter settings used. These are also motivated by the cholesterol data.

The simulation study involves simulating data from the true model (6); computing point and interval estimates and performing the test of homoscedasticity using each of the three fitting methods; repeating the entire process 500 times; and obtaining the desired estimates of the performance measures. For greater accuracy, inference on the variance components (τ^2, ψ^2, σ_1^2, and σ_2^2) is performed on log scale. The GH method uses $M = 30$ nodes.

Table 2 presents the estimated biases of point estimators in case of $n_{ij} = 2$. The biases for β_1 and μ are negligible relative to their true values for all settings. The biases are small, negative for $\log \tau^2$ and $\log \psi^2$. The situation is less clear for other parameters as the biases of their estimators may be positive or negative, albeit their magnitudes are relatively small. Moreover, there is no method that produces the smallest bias for all parameters. The same qualitative conclusions hold in case of $n_{ij} = 3$ (results not presented).

Tables 3 and 4 present estimated efficiencies of LA and MA relative to GH, defined as $\text{MSE}_{\text{LA}}/\text{MSE}_{\text{GH}}$ and $\text{MSE}_{\text{MA}}/\text{MSE}_{\text{GH}}$, respectively. We see that the efficiency depends on the parameter, the level of heteroscedasticity and the number of replications. The entries in Table 3 lie between 0.98 to 1.29, implying that GH is more accurate than LA. This finding is not unexpected because LA is a special case of GH with one node. Although there is no difference in the two methods for homoscedastic case, GH's gain in accuracy can be substantial when the extent of heteroscedasticity is high and there are three replications.

The conclusion is less clear-cut when we look at Table 4 but 90% of entries are between 0.77 to 1.10, implying that in a vast majority of cases MA produces either nearly as accurate or more accurate estimate than GH. There are a few entries greater than 1.10, but it is hard to see a simple pattern among them except that most occur when $\delta > 0$.

Table 5 presents estimated coverage probabilities of 95% confidence intervals in case of $n_{ij} = 2$. All entries are quite close to 95% when $\delta = 0$. But the performance of LA and GH methods degrades as δ increases, and it is not acceptable when $\delta \geq 1$, especially in settings 2 and 3. On the other hand, MA method maintains its coverage probability reasonably close to 95% in all cases. The results for $n_{ij} = 3$ are omitted as they lead to the same conclusion.

Table 6 presents estimated type I error probabilities for 5% level likelihood ratio test for homoscedasticity. All entries are reasonably close to 5%, implying that there is little to distinguish between the three estimation methods on this criterion.

Taken together, these findings allow us to conclude that MA is the best choice among the three model fitting methods. Not only it is simplest to implement but it also generally produces the most accurate point and interval estimates. Besides, the test of homoscedasticity based on it has type I error rates close to the nominal level. We also see that $m = 50$ subjects is large enough for acceptable accuracy of this method. The two numerical approximation methods — LA and GH — do not perform as well with $m = 50$.

5 Evaluation of similarity and agreement

Evaluation of similarity of measurement methods and agreement between them are two key goals in a method comparison study. This evaluation is conducted by performing inference on measures of similarity and agreement, which are functions of the model parameters. Now we take up the task of obtaining these measures and performing inference on them under

(6) as the data model. The task includes examining biases and precisions of methods and also the marginal and joint distributions of their measurements. These entities are easy to define and interpret when the model is linear in b. Therefore, instead of the original model (6) we work with its approximation (14) wherein b appears linearly. Further, to make the exposition simpler, we use the companion model of (14) for $\tilde{\mathbf{Y}}$. It can be written as

$$\begin{aligned}
&\tilde{Y}_1 = b + b_1 + e_1,\ \tilde{Y}_2 \approx d(b^*, \boldsymbol{\beta}) + f'(b^*, \boldsymbol{\beta})b + b_2 + e_2; \\
&b_j \sim \mathcal{N}_1(0, \psi^2),\ e_j \dot{\sim} \mathcal{N}_1\big(0, \sigma_j^2\, g_j^2(b^*, \boldsymbol{\delta}_j)\big),\ b \sim \mathcal{N}_1(\mu, \tau^2),\ j = 1, 2,
\end{aligned} \tag{18}$$

where d is defined in (15) and b^* — a fixed quantity close to b — serves as a proxy for the magnitude of measurement. As before, by marginalizing over (b, b_1, b_2) we see that $\tilde{\mathbf{Y}} \dot{\sim} \mathcal{N}_2\big(E[\tilde{\mathbf{Y}}], \mathrm{var}[\tilde{\mathbf{Y}}]\big)$ with

$$E[\tilde{\mathbf{Y}}] \approx \begin{bmatrix} \mu \\ d(b^*, \boldsymbol{\beta}) + f'(b^*, \boldsymbol{\beta})\mu \end{bmatrix} \tag{19}$$

and

$$\mathrm{var}(\tilde{\mathbf{Y}}) \approx \begin{bmatrix} \tau^2 + \psi^2 + \sigma_1^2\, g_1^2(b^*, \boldsymbol{\delta}_1) & \tau^2 f'(b^*, \boldsymbol{\beta}) \\ \tau^2 f'(b^*, \boldsymbol{\beta}) & \tau^2\{f'(b^*, \boldsymbol{\beta})\}^2 + \psi^2 + \sigma_2^2\, g_2^2(b^*, \boldsymbol{\delta}_2) \end{bmatrix}. \tag{20}$$

For the difference $\tilde{D} = \tilde{Y}_1 - \tilde{Y}_2$, it follows that $\tilde{D} \dot{\sim} \mathcal{N}_1\big(E[\tilde{D}], \mathrm{var}[\tilde{D}]\big)$, where

$$\begin{aligned}
&E[\tilde{D}] \approx -d(b^*, \boldsymbol{\beta}) + \big(1 - f'(b^*, \boldsymbol{\beta})\big)\mu, \\
&\mathrm{var}[\tilde{D}] \approx \tau^2\big(1 - f'(b^*, \boldsymbol{\beta})\big)^2 + 2\psi^2 + \sigma_1^2\, g_1^2(b^*, \boldsymbol{\delta}_1) + \sigma_2^2\, g_2^2(b^*, \boldsymbol{\delta}_2).
\end{aligned} \tag{21}$$

Both the distributions depend on $b^* \in \mathcal{B}$, which we take as the observed range of the data.

5.1 Measures of similarity

Measures of similarity compare features of marginal distributions of methods, such as biases and precisions. From models (1) and (18) for $\tilde{Y}_2$ we see that the intercept $d(b^*, \boldsymbol{\beta})$ and the

slope $f'(b^*, \boldsymbol{\beta})$ can be respectively interpreted as the "fixed" bias and the "proportional" bias of the test method. These biases depend on b^* unless $f(b, \boldsymbol{\beta}) = \beta_0 + \beta_1 b$, in which case $d(b^*, \boldsymbol{\beta}) = \beta_0$ and $f'(b^*, \boldsymbol{\beta}) = \beta_1$. If the slope is one, the methods have the same scale. If, in addition, the intercept is also zero, the methods have the same true values.

Precisions of methods can be compared via their ratio but it requires the methods to have the same scale [9, pp. 49-50]. The scale of the test method can be made same as the reference method by dividing $\tilde{Y}_2$ by the slope $f'(b^*, \boldsymbol{\beta})$. The precision ratio then becomes

$$\lambda(b^*) = \{f'(b^*, \boldsymbol{\beta})\}^2 \frac{\sigma_1^2 \, g_1^2(b^*, \boldsymbol{\delta}_1)}{\sigma_2^2 \, g_2^2(b^*, \boldsymbol{\delta}_2)}, \quad b^* \in \mathcal{B}. \tag{22}$$

This ratio depends on b^* unless the model is homoscedastic and f is linear in b. If $\lambda(b^*) < 1$, the reference method is more precise at b^* than the test method, and vice versa. While this ratio compares $\text{var}[e_j]$, often a comparison of $\text{var}[b_j + e_j]$ is of interest [9, p. 115]. This can be done by replacing $\sigma_j^2 g_j^2(b^*, \boldsymbol{\delta}_j)$ in (22) with $\psi^2 + \sigma_j^2 g_j^2(b^*, \boldsymbol{\delta}_j)$, $j = 1, 2$.

5.2 Measures of agreement

In contrast to the measures of similarity that compare marginal distributions of methods, the measures of agreement essentially look at their joint distribution to quantify how close the individual measurements are. The methods agree perfectly well if their measurements are identical. Potentially one can directly evaluate agreement between $(\tilde{Y}_1, \tilde{Y}_2)$. In this case, the effect of any fixed or proportional bias that may exist in the test method manifests in the agreement measures, which are functions of parameters of the bivariate distribution of $(\tilde{Y}_1, \tilde{Y}_2)$. However, this approach is appropriate only if the similarity evaluation does not show any proportional bias in the test method because in this case the methods are on the same scale, and hence are comparable. Otherwise, it is more appropriate to recalibrate the test method to make its scale same as the reference method prior to evaluating agreement.

This is just like the rescaling done in (22) before comparing precisions. Taking the rescaling a step further, one can additionally remove any fixed bias that may be present in the test method besides the proportional bias by transforming its measurements as

$$\tilde{Y}_2^* = \frac{\tilde{Y}_2 - d(b^*, \boldsymbol{\beta})}{f'(b^*, \boldsymbol{\beta})}. \tag{23}$$

The recalibrated test method has the same true value as the reference method. It follows from (18)-(20) that

$$\begin{bmatrix} \tilde{Y}_1 \\ \tilde{Y}_2^* \end{bmatrix} \dot{\sim} \mathcal{N}_2 \left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \tau^2 + \psi^2 + \sigma_1^2\, g_1^2(b^*, \boldsymbol{\delta}_1) & \tau^2 \\ \tau^2 & \tau^2 + \frac{\psi^2 + \sigma_2^2\, g_2^2(b^*, \boldsymbol{\delta}_2)}{\{f'(b^*, \boldsymbol{\beta})\}^2} \end{bmatrix} \right). \tag{24}$$

Moreover, the difference $\tilde{D}^* = \tilde{Y}_1 - \tilde{Y}_2^* \dot{\sim} \mathcal{N}_1\big(0, \text{var}[\tilde{D}^*]\big)$, where

$$\text{var}[\tilde{D}^*] \approx \psi^2 + \sigma_1^2\, g_1^2(b^*, \boldsymbol{\delta}_1) + \frac{\psi^2 + \sigma_2^2\, g_2^2(b^*, \boldsymbol{\delta}_2)}{\{f'(b^*, \boldsymbol{\beta})\}^2}. \tag{25}$$

The measures of agreement in this case are functions of parameters of the bivariate distribution of $(\tilde{Y}_1, \tilde{Y}_2^*)$. While it is true that the recalibration is likely to make the test method agree more with the reference method, but measuring their agreement is appropriate in the first place only if the two methods are on the same scale.

The expression for any measure of agreement between either $(\tilde{Y}_1, \tilde{Y}_2)$ or $(\tilde{Y}_1, \tilde{Y}_2^*)$ can be obtained by simply taking the definition of the measure and plugging-in the relevant parameters from their respective bivariate distributions. This approach works for any measure of agreement available in the literature. For example, the two versions of the agreement measure *concordance correlation coefficient* (CCC) [2] are

$$\text{CCC}(b^*) \approx \frac{2\text{cov}[\tilde{Y}_1, \tilde{Y}_2]}{\{E[\tilde{D}]\}^2 + \text{var}[\tilde{Y}_1] + \text{var}[\tilde{Y}_2]},\ \text{CCC}^*(b^*) \approx \frac{2\text{cov}[\tilde{Y}_1, \tilde{Y}_2^*]}{\text{var}[\tilde{Y}_1] + \text{var}[\tilde{Y}_2^*]},\quad b^* \in \mathcal{B},$$

where the moments are from (20), (21) and (24). The CCC lies in $[-1, 1]$ and the larger positive its value the better is the agreement. The starred version of the measure is for the recalibrated data.

The *total deviation index* (TDI) [3, 4] is another agreement measure. It is defined as the $100p$th percentile of absolute difference in measurements, where p is a specified large probability, typically between 0.80 and 0.95. The two versions of TDI can be written as

$$\begin{aligned}
&\text{TDI}(b^*, p) = 100p\text{th percentile of } |\tilde{D}| \approx \text{sd}[\tilde{D}]\big\{\chi_1^2\big(p, \big(E[\tilde{D}]/\text{sd}[\tilde{D}]\big)^2\big)\big\}^{1/2},\\
&\text{TDI}^*(b^*, p) = 100p\text{th percentile of } |\tilde{D}^*| \approx \text{sd}[\tilde{D}^*]\big\{\chi_1^2(p, 0)\big\}^{1/2}, \quad b^* \in \mathcal{B},
\end{aligned} \tag{26}$$

where the moments are from (21) and (25), and $\chi_1^2(p, \Delta)$ denotes the $100p$th percentile of a noncentral chi-squared distributed with one degree of freedom and noncentrality parameter Δ. When the noncentrality parameter is zero, $\{\chi_1^2(p, 0)\}^{1/2} = z\big((1+p)/2\big)$, the $100(1+p)$th percentile of a standard normal distribution. The TDI is a non-negative measure and the smaller its value the better is the agreement. We only focus on TDI with $p = 0.90$ for the illustration here.

5.3 Inference on measures of similarity and agreement

All measures of similarity and agreement are functions of $\boldsymbol{\theta}$ and b^*. Let ϕ denote any such measure and $\phi(b^*)$ be its value at $b^* \in \mathcal{B}$. The measure is assumed to be a scalar quantity. Replacing $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$ in its expression gives its ML estimator $\hat{\phi}(b^*)$. From delta method [31], when m is large, $\hat{\phi}(b^*) \dot{\sim} \mathcal{N}_1\big(\phi(b^*), \mathbf{G}'(b^*)\mathbf{I}^{-1}\mathbf{G}(b^*)\big)$, where $\mathbf{G}(b^*) = (\partial/\partial\boldsymbol{\theta})\phi(b^*)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ can be computed numerically. Thus, approximate $100(1-\alpha)\%$ two-sided pointwise confidence interval for $\phi(b^*)$ on a grid of values of $b^* \in \mathcal{B}$ can be computed as $\hat{\phi}(b^*) \pm z_{1-\alpha/2}\ \{\mathbf{G}'(b^*)\mathbf{I}^{-1}\mathbf{G}(b^*)\}^{1/2}$. One-sided pointwise bands for $\phi(b^*)$ can by computed by replacing $z_{1-\alpha/2}$ with $z_{1-\alpha}$ and using only either the lower limit or the upper limit of this interval. If ϕ is a measure of similarity, a two-sided confidence interval for ϕ is of interest. On the other hand, if ϕ is a measure of agreement, an appropriate one-sided confidence bound for ϕ is of interest. In particular, if small values for ϕ imply good agreement (e.g., TDI), we need an upper bound.

Whereas, if large values for ϕ imply good agreement (e.g., CCC), we need a lower bound. These bounds and intervals can be made more accurate by computed them after applying a suitable normalizing transformation to the measure (e.g., log transformation of TDI or Fisher's z-transformation for CCC) and back-transforming the results to the original scale. These confidence bounds and intervals are used to evaluate similarity and agreement over the measurement range $\mathcal{B}$.

6 Application to cholesterol data

The cholesterol data come from a trial conducted at Virginia Commonwealth University to compare two methods for assaying serum cholesterol (mg/dL) [12]. One is Cobas Bio, an assay standardized by the Centers for Disease Control that serves as the reference method (method 1). The other is Ektachem 700, a routine laboratory analyzer that serves as the test method (method 2). There are 100 subjects in this study. Measurements from each assay are replicated ten times on every subject. The replications from an assay are exchangeable and the replications across the assays are unpaired. The measurements range between 45 to 372 mg/dL. Figure 1 shows a trellis plot the data from [23]. We see that Ektachem's measurements tend to be larger and have higher within-subject variation than Cobas Bio's. This variation increases with cholesterol level for both assays. There is also evidence of assay $\times$ subject interaction. In our notation, Y_{ijk} represents the kth replicate measurement of serum cholesterol obtained by the jth assay from the blood sample of the ith subject, $i = 1, \ldots, 100$, $j = 1, 2$, and $k = 1, \ldots, 10$.

The first step in the analysis is modeling of data. This involves specifying parametric forms for the calibration function f and the variance functions g_j, $j = 1, 2$, in (6). Figure 2 shows a scatter plot of $(\overline{y}_{i1}, \overline{y}_{i2})$ pairs and also assay-specific plots of the logarithm of SD of

a subject's ten measurements against the logarithm of their mean. The points in each plot cluster around a straight line, suggesting $f(b, \boldsymbol{\beta}) = \beta_0 + \beta_1 b$ and $g_j(b, \delta_j) = |b|^{\delta_j}$, $j = 1, 2$, as plausible choices. The same choice for g_j is suggested when residuals from a homoscedastic fit are analyzed. With these f and g_j, the resulting model (6) has nine parameters. Table 7 summarizes estimates of these parameters and their SEs obtained using the three model fitting methods described in Section 3. Although the three methods produce practically the same results, there is a slight difference in the estimates of β_0 obtained by numerical approximation methods (LA and GH) and the model approximation method (MA). Nevertheless, the difference is not large enough to be of concern. Besides, the simulations in Section 4 show that the MA method is generally more accurate than the other two methods anyway. Therefore, only the results from MA method will be presented hereafter. Figure 2 (d) shows a plot of standardized residuals against the fitted values. It has no discernible pattern. This, together with additional model diagnostics suggested in [10] (not presented here), allows us to conclude that the fit of the assumed model is adequate.

The p-value for the likelihood ratio test of null hypothesis of homoscedasticity is practically zero, confirming nonconstant error variances. Plugging-in parameter estimates in (18)-(20) gives the fitted distribution of $(\tilde{Y}_1, \tilde{Y}_2)$ as

$$\begin{bmatrix} \tilde{Y}_1 \\ \tilde{Y}_2 \end{bmatrix} \dot{\sim} \mathcal{N}_2 \left(\begin{bmatrix} 184.38 \\ 190.24 \end{bmatrix}, \begin{bmatrix} 4255.97 + (8.0 \times 10^{-5}) b^{*2.04} & 4314.78 \\ 4314.78 & 4426.87 + (1.9 \times 10^{-4}) b^{*1.98} \end{bmatrix} \right).$$

This distribution depends on the cholesterol level $b^* \in \mathcal{B} = [45, 372]$ mg/dL because of heteroscedasticity. Notice that the contribution of error variation to the total variation in response is swamped by other components of variation. In particular, this makes the estimated correlation between $(\tilde{Y}_1, \tilde{Y}_2)$ very high — over 0.985 — throughout $\mathcal{B}$.

The second step in the analysis is evaluation of similarity. The estimate of proportional bias β_1 is 1.02 (SE = 0.01) and its 95% confidence interval is $[1.00, 1.04]$. Thus, there

is evidence of a slight upward proportional bias of up to 4% in Ektachem assay, but the evidence is borderline. Further, the estimate of fixed bias β_0 is 2.17 (SE = 2.20) and its 95% confidence interval is $[-2.14, 6.48]$. Although this interval covers zero, it also provides evidence of a small fixed bias. These findings are consistent with the observation that Ektachem's measurements tend to be larger than Cobas Bio's. Figure 3 presents estimate and 95% two-sided pointwise confidence band for precision ratio λ, defined in (22), as a function of cholesterol level b^*. The entire band lies below one. Notwithstanding the fact this band is pointwise rather than simultaneous, it does indicate that Cobas Bio is more precise than Ektachem. The former is estimated to be about 40% more precise than the latter. To summarize, we find that the two assays cannot be regarded as similar. Not only they do not have the same true values, but also Cobas Bio is more precise than Ektachem.

The third step in the analysis is evaluation of agreement. Since there is evidence of a slight bias in Ektachem, we use (23) to recalibrate its measurement $\tilde{Y}_2$ as $\tilde{Y}_2^*$ to make its true value same as Cobas Bio's. The estimated transformation is $\tilde{Y}_2^* = (\tilde{Y}_2 - 2.17)/1.02$. Using (25), the fitted distribution of the difference $\tilde{D}^*$ after the transformation is

$$\tilde{D}^* \dot{\sim} \mathcal{N}_1\left(0, 50.58 + (8.0 \times 10^{-5}){b^*}^{2.04} + (1.8 \times 10^{-4}){b^*}^{1.98}\right).$$

The SD of this distribution ranges between 7.15 to 9.31. Next, we perform inference on the agreement measure TDI* (with $p = 0.90$), given by (26), as described in Section 5. Figure 3 shows its 95% pointwise upper confidence bound as a function of cholesterol level b^*. This bound increases from 13.6 to 16.7 as the cholesterol level increases from 45 to 372 mg/dL. The leftmost bound of 13.6 shows that 90% of differences in measurements from the assays when the true value is 45 fall within ± 13.6. Relative to the true value, this difference is too large to be deemed acceptable. On the other hand, the rightmost bound of 16.7 shows that 90% of differences in measurements from the assays when the true value is 372 fall within ± 16.7. This difference may be considered acceptable. Thus, we may conclude that the

assays, after the recalibration, have satisfactory agreement for large cholesterol values but not for small values. Obviously, this means that the Cobas Bio and recalibrated Ektachem do not agree well enough to be considered interchangeable. In fact, we know from evaluation of similarity that Cobas Bio is superior to Ektachem by virtue of being more precise. It may be noted that if Ektachem is not recalibrated prior to agreement evaluation, then the 95% pointwise upper confidence bound for TDI ranges from 17.2 to 19.8 over $\mathcal{B}$. These bounds are a bit larger than before because of Ektachem's bias, and hence imply a somewhat worse level of agreement between the two assays.

To see the effect of ignoring heteroscedasticity, we repeat the analysis assuming constant error variances, i.e., setting the heteroscedasticity parameters in (6) to zero. The estimate of TDI^* and its 95% confidence bound come out to be 12.9 and 14.5, respectively. Although these quantities do not depend on the cholesterol value, they are not too far from their heteroscedastic counterparts that range between 11.8 to 15.3 and 13.6 to 16.7, respectively. This happens because the error variation, albeit nonconstant, is swamped by other variance components that do not change with cholesterol value. Nevertheless, it is apparent that the homoscedastic model underestimates the extent of agreement for small cholesterol values and overestimates it for large cholesterol values.

7 Discussion

This article presents a measurement error model for replicated method comparison data that can incorporate heteroscedasticity of errors as well as nonlinear relationships in the true values of the measurement methods. It also shows how the model can be used to evaluate similarity and agreement between the methods. A key advantage of the model is that it allows one method to be recalibrated against the other, either linearly or nonlinearly, to ensure that

their true values are identical. Here we focussed on comparison of two methods and did not include covariates. But the model can be extended to accommodate more than two methods and covariates. We also assumed normality for random effects and error distributions. The model can deal with skewness and heavytailedness in the data by replacing the normality assumption with generalizations of normal, such as skew-normal and skew-t distributions. We, however, require the measurements to be replicated to avoid identifiability issues. The model also requires the practitioner to specify parametric forms for calibration and variance functions, as one ordinarily does in regression modeling. Further research is needed to allow these functions to be specified semiparametrically or even nonparametrically.

A potential limitation of our model (6) or its linearized version (14) is that the interaction effects of the two methods have the same variance even though the methods may have different scales. Without the equal variance assumption, the model is not identifiable in the linear calibration case. If this assumption is a concern, it can be addressed to some extent by replacing the interaction effect b_{i2} of the test method in (14) with $f'(b_i^*, \boldsymbol{\beta})\, b_{i2}$, making the new effect's variance different from that of the reference method. In the linear calibration case, this means b_{i2} is replaced with $\beta_1 b_{i2}$. The change in (14) can be easily propagated through subsequent steps of the analysis to get the analysis based on the new model.

Acknowledgements

The authors thank Professor Subharup Guha for asking questions that spurred this work. They also thank the reviewers for their constructive comments. Thanks are also due to Professor Vernon Chinchilli for providing the cholesterol data, and the Texas Advanced Computing Center at The University of Texas at Austin for providing HPC resources for conducting the simulation studies.

References

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

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

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

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

5. Haber MJ, Barnhart HX. A general approach to evaluating agreement between two observers or methods of measurement from quantitative data with replicated measurements. *Statistical Methods in Medical Research* 2008; **17**:151–169.

6. Pan Y, Haber M, Gao J, Barnhart HX. A new permutation-based method for assessing agreement between observers making replicated quantitative readings. *Statistics in Medicine* 2012; **31**:2249–2261.

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

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

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

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

11. Cheng C, Van Ness JW. *Statistical Regression with Measurement Error*. John Wiley, Chichester, UK, 1999.

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

13. 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.

14. 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.

15. 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.

16. 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.

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

18. Dunn G, Roberts C. Modelling method comparison data. *Statistical Methods in Medical Research* 1999; **8**:161–179.

19. Carstensen B. Comparing methods of measurement: Extending the LoA by regression. *Statistics in Medicine* 2010b; **29**:401–410.

20. Alanen E. Everything all right in method comparison studies? *Statistical Methods in Medical Research* 2012; **21**:297–309.

21. Kelly, G. E. Use of structural equations model in assessing the reliability of a new measurement technique. *Applied Statistics* 1985; **34**:258–263.

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

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

24. Linnet K. Estimation of the linear relationship between the measurements of two methods with proportional errors. *Statistics in Medicine* 1990; **9**:1463–1473.

25. Meijer E, Mooijaart A. Factor analysis with heteroscedastic errors. *British Journal of Mathematical and Statistical Psychology* 1996; **49**:189–202.

26. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. `http://www.R-project.org`.

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

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

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

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

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

Table 1: Sets of parameter values used for the simulation study.

	Set		
$\boldsymbol{\theta}$	1	2	3
(β_0, β_1)	(10, 1.2)	(5, 1.1)	(0, 1)
$(\mu, \log(\tau^2), \log(\psi^2))$	(185, 8, 3)	(185, 8, 3)	(185, 8, 3)
$(\log(\sigma_1^2), \log(\sigma_2^2))$	*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 2: Estimated biases of estimators computed using three model fitting methods.

			$\delta = 0$			$\delta = 0.9$			$\delta = 1$			$\delta = 1.1$	
Set	$\boldsymbol{\theta}$	LA	GH	MA	LA	GH	MA	LA	GH	MA	LA	GH	MA
1	β_0	0.24	0.24	0.02	-0.10	-0.02	-0.01	-0.05	-0.04	0.13	-0.22	-0.18	-0.27
	β_1	0.00	0.00	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00
	μ	0.09	0.11	-0.24	0.78	0.69	0.50	2.13	2.12	0.02	2.27	1.91	-0.01
	$\log \tau^2$	-0.04	-0.04	-0.04	-0.07	-0.07	-0.04	-0.09	-0.09	-0.03	-0.12	-0.12	-0.04
	$\log \psi^2$	-0.07	-0.07	-0.07	-0.04	-0.05	-0.07	-0.07	-0.07	-0.07	-0.15	-0.14	-0.11
	$\log \sigma_1^2$	-0.14	-0.14	-0.28	-0.08	-0.10	0.07	-0.19	-0.15	-0.28	0.03	0.08	-0.12
	$\log \sigma_2^2$	-0.50	-0.50	-0.46	0.50	0.49	-0.52	0.57	0.58	-0.25	0.11	0.11	-0.19
	δ_1	0.01	0.01	0.02	0.00	0.01	-0.01	0.01	0.01	0.02	-0.01	-0.01	0.01
	δ_2	0.04	0.04	0.04	-0.05	-0.05	0.04	-0.06	-0.06	0.02	-0.01	-0.01	0.02
2	β_0	-0.31	-0.31	-0.07	0.12	0.10	-0.03	0.05	0.05	0.09	-0.36	-0.43	-0.20
	β_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	μ	-0.04	-0.05	0.03	0.80	0.56	-0.08	1.74	1.67	0.09	2.36	2.34	0.16
	$\log \tau^2$	-0.06	-0.06	-0.03	-0.05	-0.04	-0.04	-0.10	-0.09	-0.02	-0.12	-0.12	-0.05
	$\log \psi^2$	-0.06	-0.06	-0.07	-0.03	-0.03	-0.04	-0.05	-0.05	-0.10	-0.09	-0.09	-0.10
	$\log \sigma_1^2$	-0.15	-0.17	-0.32	0.04	0.03	-0.27	0.19	0.18	-0.22	0.14	0.17	0.04
	$\log \sigma_2^2$	-0.40	-0.39	-0.25	0.22	0.18	-0.33	0.41	0.39	-0.36	0.05	-0.04	-0.19
	δ_1	0.01	0.01	0.03	-0.01	-0.01	0.02	-0.02	-0.02	0.02	-0.02	-0.02	-0.01
	δ_2	0.03	0.03	0.02	-0.03	-0.02	0.03	-0.04	-0.04	0.03	-0.01	0.00	0.01
3	β_0	-0.03	-0.02	-0.11	-0.17	-0.15	-0.22	-0.26	-0.29	-0.05	-0.20	-0.21	-0.35
	β_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	μ	-0.35	-0.40	-0.32	0.03	-0.11	-0.06	1.89	1.74	0.04	2.52	2.49	-0.14
	$\log \tau^2$	-0.03	-0.03	-0.05	-0.05	-0.04	-0.03	-0.08	-0.08	-0.04	-0.12	-0.12	-0.02
	$\log \psi^2$	-0.07	-0.07	-0.07	-0.05	-0.05	-0.05	-0.07	-0.07	-0.07	-0.09	-0.08	-0.07
	$\log \sigma_1^2$	-0.40	-0.41	-0.29	-0.09	-0.05	-0.33	0.17	0.15	-0.25	-0.14	-0.11	-0.29
	$\log \sigma_2^2$	-0.13	-0.13	-0.24	0.11	0.14	-0.34	0.12	0.16	-0.49	0.16	0.15	-0.19
	δ_1	0.03	0.04	0.02	0.00	0.00	0.03	-0.02	-0.02	0.02	0.01	0.01	0.03
	δ_2	0.01	0.01	0.02	-0.01	-0.02	0.03	-0.02	-0.02	0.04	-0.02	-0.02	0.02

Table 3: Relative efficiencies ($\text{MSE}_{\text{LA}}/\text{MSE}_{\text{GH}}$) of estimates obtained by LA and GH methods.

			$\boldsymbol{\theta}$								
δ	n_{ij}	Set	β_0	β_1	μ	$\log\tau^2$	$\log\psi^2$	$\log\sigma_1^2$	$\log\sigma_2^2$	δ_1	δ_2
0	2	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		2	1.00	1.00	1.00	1.00	1.00	1.01	1.00	1.02	1.00
		3	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00
	3	1	1.00	1.00	1.01	1.00	1.00	1.00	1.00	1.00	1.00
		2	1.00	1.00	1.00	1.00	1.00	0.99	1.01	1.00	1.01
		3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.9	2	1	1.08	1.07	1.02	1.01	1.00	1.04	1.02	1.04	1.02
		2	1.06	1.05	1.02	1.00	1.01	1.04	1.05	1.04	1.05
		3	1.08	1.08	1.02	1.04	1.05	1.09	1.06	1.09	1.06
	3	1	1.14	1.14	1.14	1.01	1.06	1.18	1.08	1.18	1.09
		2	1.21	1.21	1.24	1.20	1.18	1.29	1.25	1.29	1.25
		3	1.14	1.13	1.10	1.14	1.10	1.21	1.13	1.20	1.13
1	2	1	1.03	1.03	1.00	1.01	1.02	1.11	1.05	1.10	1.05
		2	1.01	1.01	1.04	0.98	1.01	1.01	0.99	1.01	1.00
		3	1.03	1.03	1.02	1.04	1.02	1.02	1.08	1.02	1.08
	3	1	1.19	1.20	1.25	1.05	1.07	1.24	1.21	1.24	1.21
		2	1.21	1.20	1.14	1.10	1.02	1.21	1.18	1.21	1.18
		3	1.22	1.18	1.23	1.11	1.11	1.12	1.18	1.13	1.18
1.1	2	1	1.12	1.10	1.12	1.07	1.06	1.11	1.13	1.11	1.13
		2	1.06	1.04	1.06	1.01	0.98	1.09	1.04	1.09	1.04
		3	1.05	1.05	1.03	1.01	1.03	1.06	1.03	1.06	1.03
	3	1	1.19	1.16	1.21	1.11	1.10	1.19	1.15	1.19	1.15
		2	1.15	1.15	1.26	1.06	1.02	1.22	1.18	1.21	1.19
		3	1.07	1.06	1.05	1.03	1.04	1.12	1.16	1.12	1.16

Table 4: Relative efficiencies ($\mathrm{MSE_{MA}}/\mathrm{MSE_{GH}}$) of estimates obtained by MA and GH methods.

			$\boldsymbol{\theta}$								
δ	n_{ij}	Set	β_0	β_1	μ	$\log\tau^2$	$\log\psi^2$	$\log\sigma_1^2$	$\log\sigma_2^2$	δ_1	δ_2
0	2	1	0.99	1.00	0.95	1.00	1.00	0.86	1.01	0.86	1.02
		2	0.99	0.92	0.97	1.00	1.07	1.07	0.84	1.07	0.85
		3	1.12	1.16	1.02	1.11	1.02	0.93	0.91	0.95	0.91
	3	1	0.91	0.92	0.94	0.95	0.98	0.93	0.93	0.93	0.93
		2	0.94	0.94	0.96	0.96	0.97	0.92	0.92	0.92	0.91
		3	0.93	0.94	0.95	0.97	0.98	0.92	0.95	0.92	0.95
0.9	2	1	0.96	0.91	1.18	0.95	0.97	1.09	0.95	1.10	0.96
		2	1.04	0.99	0.83	1.05	1.02	1.16	0.91	1.18	0.91
		3	1.02	0.97	0.98	0.96	1.22	1.13	1.09	1.12	1.09
	3	1	1.00	0.99	0.95	0.96	0.92	0.93	0.91	0.92	0.91
		2	1.00	0.98	1.06	1.07	1.07	1.07	1.01	1.06	1.01
		3	0.97	0.96	0.96	1.01	0.96	1.01	0.98	1.01	0.98
1	2	1	1.11	1.11	1.19	0.84	1.22	1.11	1.03	1.11	1.03
		2	0.94	0.98	1.14	0.82	1.15	0.85	0.82	0.85	0.82
		3	0.90	0.88	0.89	0.99	1.00	1.00	1.01	0.99	1.01
	3	1	1.09	1.10	1.10	1.02	1.11	1.01	1.07	1.01	1.07
		2	1.01	1.03	1.00	1.02	1.05	0.99	1.04	1.00	1.04
		3	1.04	1.02	1.07	1.02	1.03	0.95	0.94	0.95	0.94
1.1	2	1	0.81	0.86	0.99	1.03	0.81	1.01	0.90	1.02	0.89
		2	1.00	0.98	1.04	0.83	1.05	0.80	0.83	0.80	0.86
		3	1.01	1.00	1.02	0.77	0.97	1.03	0.97	1.04	0.98
	3	1	1.06	1.04	1.06	1.07	1.00	1.04	1.00	1.05	1.01
		2	1.09	1.08	1.27	1.07	1.01	1.08	1.07	1.07	1.08
		3	1.12	1.13	1.18	1.07	0.99	1.06	1.06	1.06	1.06

Table 5: Estimated coverage probabilities (in %) of 95% confidence intervals computed using three model fitting methods in case of $n_{ij} = 2$.

		$\delta = 0$			$\delta = 0.9$			$\delta = 1$			$\delta = 1.1$		
Set	$\boldsymbol{\theta}$	LA	GH	MA	LA	GH	MA	LA	GH	MA	LA	GH	MA
1	β_0	94.6	94.6	93.6	92.0	92.4	94.6	87.4	86.6	95.0	91.2	92.0	94.8
	β_1	94.2	94.2	93.6	92.0	92.4	95.2	78.0	79.2	94.6	91.4	92.2	95.6
	μ	95.4	95.4	94.8	95.2	95.2	95.2	90.6	90.6	96.0	93.0	93.4	94.6
	$\log \tau^2$	92.8	92.8	93.0	91.0	91.2	93.6	80.4	80.4	94.8	91.6	92.4	93.8
	$\log \psi^2$	94.2	94.2	94.2	92.4	92.6	94.2	81.2	81.6	93.8	94.2	94.6	96.4
	$\log \sigma_1^2$	95.4	96.0	94.2	94.0	94.8	93.6	84.8	84.0	95.8	94.6	96.0	95.6
	$\log \sigma_2^2$	94.6	95.6	94.8	94.4	94.6	94.6	81.4	80.8	94.6	91.4	93.8	94.4
	δ_1	95.2	95.8	94.8	92.6	93.6	93.6	75.0	76.0	95.8	95.0	95.8	95.4
	δ_2	94.8	95.6	95.0	93.0	93.2	94.8	73.8	73.2	94.6	91.0	93.0	94.4
2	β_0	94.2	95.0	93.2	93.6	95.8	95.4	88.6	89.4	95.0	87.4	90.6	92.4
	β_1	94.2	95.0	94.4	93.4	95.0	94.4	84.8	87.2	94.6	86.2	89.4	93.6
	μ	92.0	92.0	94.0	93.6	93.4	95.6	93.4	93.6	95.2	89.6	91.0	96.4
	$\log \tau^2$	93.8	93.8	95.0	93.8	93.0	92.6	88.6	87.8	95.8	83.8	85.0	93.4
	$\log \psi^2$	93.8	93.8	94.0	93.0	92.6	93.6	87.6	86.6	93.8	89.0	90.0	94.8
	$\log \sigma_1^2$	94.6	95.2	95.4	95.6	95.4	91.8	88.4	86.6	94.2	86.8	89.6	95.4
	$\log \sigma_2^2$	94.0	94.4	94.6	94.4	94.8	94.2	86.6	85.0	93.6	88.6	89.4	95.8
	δ_1	96.0	96.8	95.4	94.4	94.8	91.8	84.2	83.6	94.0	85.4	87.4	95.2
	δ_2	94.0	94.4	94.4	93.2	93.8	94.6	85.6	83.8	93.4	87.8	88.2	96.0
3	β_0	94.0	94.4	92.6	93.6	94.2	96.4	87.0	87.4	95.4	88.8	87.6	93.8
	β_1	95.0	95.4	93.0	92.6	93.8	95.8	87.4	88.6	95.2	88.0	88.8	94.4
	μ	93.4	93.4	92.2	95.0	94.8	95.4	91.6	91.2	95.0	91.6	91.4	93.8
	$\log \tau^2$	94.4	94.6	93.2	95.0	95.0	95.8	87.2	87.8	94.4	84.8	85.2	95.6
	$\log \psi^2$	92.0	92.4	92.6	94.6	94.8	92.4	85.6	85.4	93.6	89.0	89.4	95.2
	$\log \sigma_1^2$	94.4	94.8	96.2	95.0	94.8	94.2	89.2	89.2	95.0	88.4	88.8	95.2
	$\log \sigma_2^2$	93.8	94.6	95.8	95.8	96.0	95.2	87.2	88.4	94.0	87.2	88.4	95.4
	δ_1	94.6	95.0	95.8	95.0	95.2	94.6	86.2	86.2	95.8	85.6	86.0	95.0
	δ_2	93.6	94.2	95.8	95.6	96.4	95.2	83.4	85.4	94.0	83.8	84.4	95.2

Table 6: Estimated type I error probabilities (in %) for 5% level likelihood ratio test of homoscedasticity performed using three model fitting methods.

	LA		GH		MA	
Set	$n_{ij}=2$	$n_{ij}=3$	$n_{ij}=2$	$n_{ij}=3$	$n_{ij}=2$	$n_{ij}=3$
1	3.6	4.6	3.6	4.6	4.4	4.6
2	4.0	5.0	4.0	4.4	5.4	4.2
3	6.2	4.0	5.6	4.0	5.6	3.8

Table 7: Estimates of parameters and their standard errors (SEs) for cholesterol data computed using three model fitting methods.

	LA		GH		MA	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
β_0	1.99	2.20	1.97	2.21	2.17	2.20
β_1	1.02	0.01	1.02	0.01	1.02	0.01
μ	184.50	6.53	184.41	6.53	184.38	6.54
$\log \tau^2$	8.35	0.14	8.35	0.14	8.35	0.14
$\log \psi^2$	3.27	0.15	3.27	0.15	3.25	0.14
$\log \sigma_1^2$	-9.50	0.60	-9.50	0.60	-9.43	0.57
$\log \sigma_2^2$	-8.51	0.61	-8.52	0.61	-8.57	0.59
δ_1	1.02	0.06	1.02	0.06	1.02	0.06
δ_2	0.98	0.06	0.98	0.06	0.99	0.06

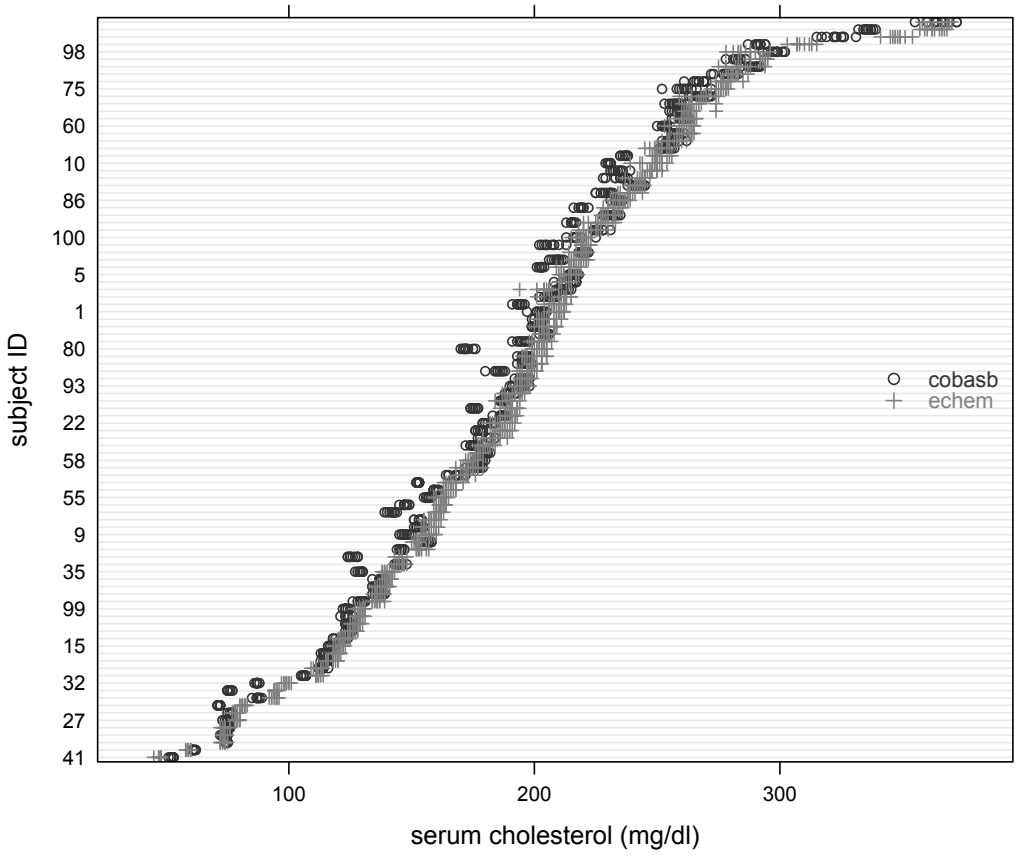

Figure 1: A trellis plot of cholesterol measurements from Cobas Bio and Ektachem assays.

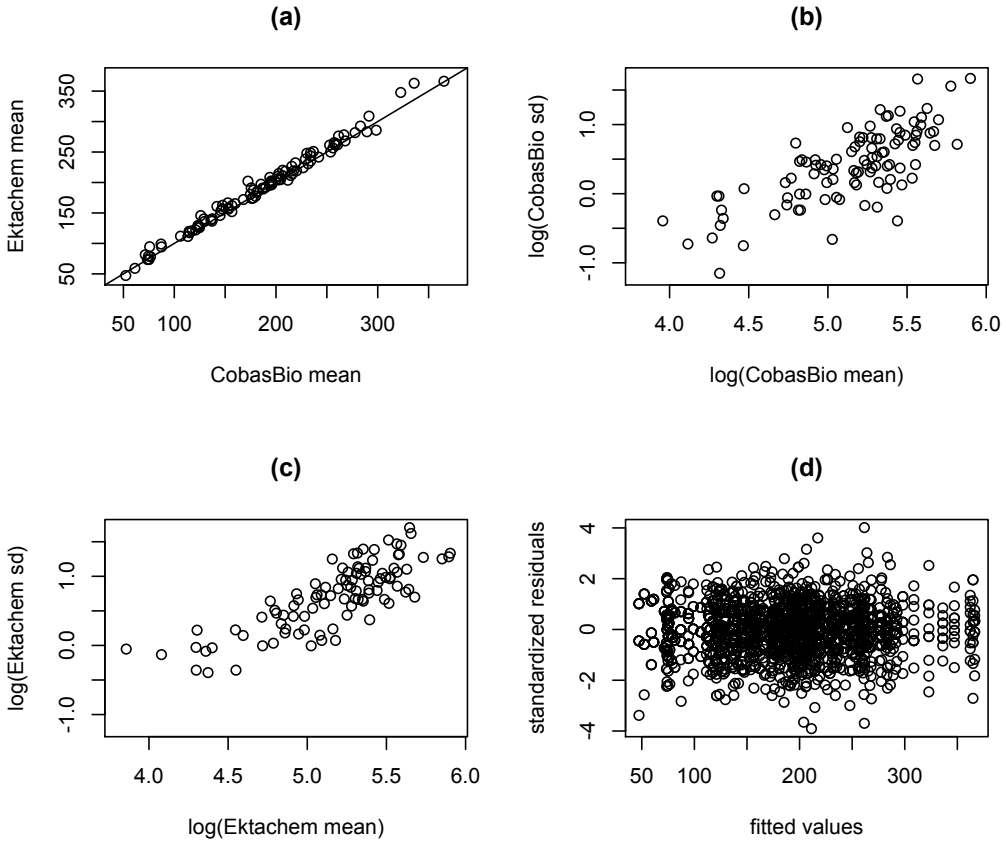

Figure 2: (a) Scatterplot of paired within-subject means with 45° degree line superimposed; (b-c) plots of within-subject SD versus within-subject mean on logarithmic scale; and (d) residual plot for cholesterol data.

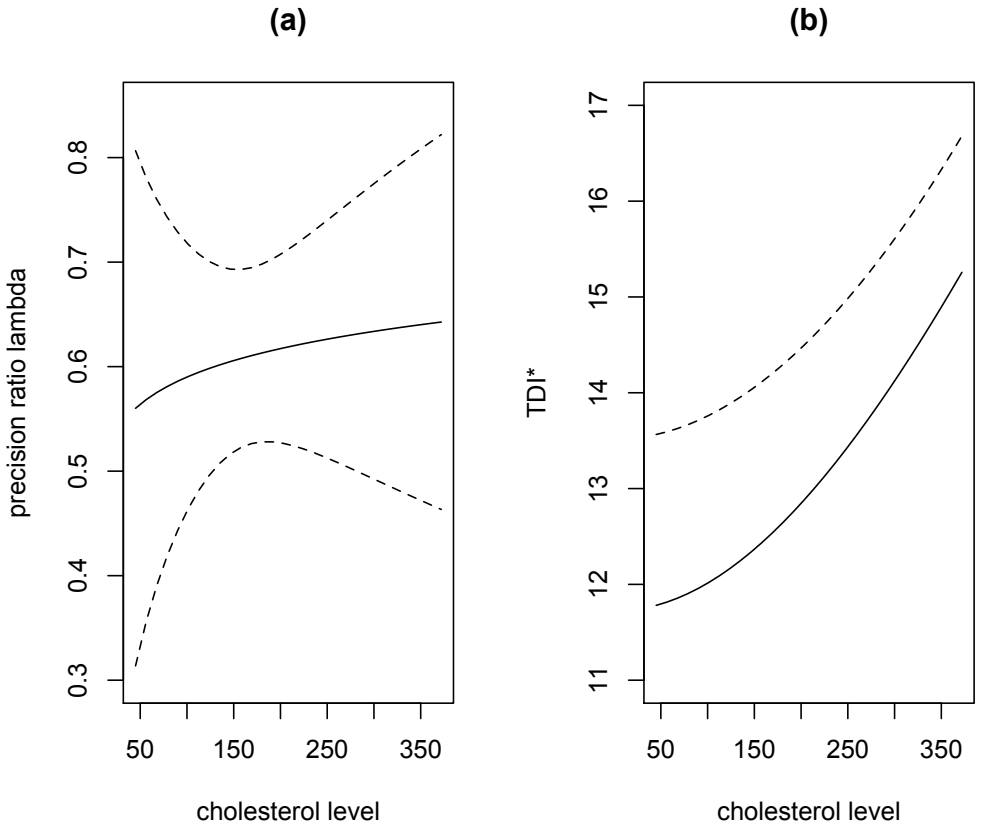

Figure 3: (a) Estimate (solid line) and 95% pointwise two-sided confidence band (broken lines) for precision ratio λ; and (b) estimate (solid line) and 95% pointwise upper confidence bound (broken line) for agreement measure TDI* with $p = 0.90$.