

# Contents

<b>1</b>	<b>LME models in method comparison studies</b>	<b>3</b>
1.1	Carstensen's Model . . . . .	4
<b>2</b>	<b>Statement of the LME model</b>	<b>5</b>
2.1	Roy's methodology . . . . .	6
<b>3</b>	<b>Agreement Criteria</b>	<b>7</b>
3.0.1	Inter-Method Bias . . . . .	9
<b>4</b>	<b>Roy's Hypotheses Tests</b>	<b>11</b>
4.1	Model Specification . . . . .	15
4.1.1	Model Terms (Roy 2009) . . . . .	15
4.2	Differences Between Approaches : Assumptions on Variability . . . . .	17
4.3	Limits of Agreement in LME models . . . . .	19
4.4	Remarks . . . . .	20
<b>5</b>	<b>Roy's Hypothesis Testing</b>	<b>20</b>
<b>6</b>	<b>Computation of limits of agreement under Roy's model</b>	<b>22</b>
<b>7</b>	<b>Carstensen's Limits of agreement</b>	<b>23</b>
<b>8</b>	<b>Roy's LME methodology for assessing agreement</b>	<b>23</b>
8.1	Roy's variability tests . . . . .	24
8.2	Hypothesis Testing . . . . .	24
8.3	Variability test 1 . . . . .	25
8.4	Variability test 2 . . . . .	25
8.5	Correlation . . . . .	25
<b>9</b>	<b>Roy's LME approach</b>	<b>27</b>

<b>10 Roy's LME methodology for assessing agreement</b>	<b>29</b>
10.0.1 Sampling Scheme : Replicates . . . . .	29
10.1 Difference Variance further to Carstensen . . . . .	30
10.2 Relevance of Roy's Methodology . . . . .	31

# 1 LME models in method comparison studies

Linear mixed effects (LME) models can facilitate greater understanding of the potential causes of bias and differences in precision between two sets of measurement. Lai and Shiao (2005) view the LME Models approach as an natural expansion to the Bland-Altman method for comparing two measurement methods. Their focus is to explain lack of agreement by means of additional covariates outside the scope of the traditional method comparison problem.

Carstensen et al. (2008) remarks that modern statistical computation, such as that used for LME models, greatly improve the efficiency of calculation compared to previous ‘by-hand’ methods. In this chapter various LME approaches to method comparison studies shall be examined.

## 1.1 Carstensen's Model

Carstensen et al. (2008) also use a LME model for the purpose of comparing two methods of measurement where replicate measurements are available on each item. Their interest lies in generalizing the popular limits-of-agreement (LOA) methodology advocated by Bland and Altman (1986) to take proper cognizance of the replicate measurements. Carstensen et al. (2008) demonstrate statistical flaws with two approaches proposed by Bland and Altman (1999) for the purpose of calculating the variance of the inter-method bias when replicate measurements are available. Instead, Carstensen et al. (2008) use a fitted mixed effects model to obtain appropriate estimates for the variance of the inter-method bias. As their interest mainly lies in extending the Bland-Altman methodology, other formal tests are not considered.

Carstensen (2004) presents a model to describe the relationship between a value of measurement and its real value. The non-replicate case is considered first, as it is the context of the Bland Altman plots. This model assumes that inter-method bias is the only difference between the two methods.

A measurement  $y_{mi}$  by method  $m$  on individual  $i$  is formulated as follows;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2) \quad (1)$$

The differences are expressed as  $d_i = y_{1i} - y_{2i}$ . For the replicate case, an interaction term  $c$  is added to the model, with an associated variance component. All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \quad (2)$$

Of particular importance is terms of the model, a true value for item  $i$  ( $\mu_i$ ). The fixed effect of Roy's model comprise of an intercept term and fixed effect terms for both methods, with no reference to the true value of any individual item. A distinction can be made between the two models: Roy's model is a standard LME model, whereas Carstensen's model is a more complex additive model.

## 2 Statement of the LME model

A linear mixed effects model is a linear model that combined fixed and random effect terms formulated by Laird and Ware (1982) as follows;

$$Y_i = X_i\beta + Z_ib_i + \epsilon_i$$

- $Y_i$  is the  $n \times 1$  response vector
- $X_i$  is the  $n \times p$  Model matrix for fixed effects
- $\beta$  is the  $p \times 1$  vector of fixed effects coefficients
- $Z_i$  is the  $n \times q$  Model matrix for random effects
- $b_i$  is the  $q \times 1$  vector of random effects coefficients, sometimes denoted as  $u_i$
- $\epsilon$  is the  $n \times 1$  vector of observation errors

## 2.1 Roy's methodology

Barnhart et al. (2007) describes the sources of disagreement as differing population means, different between-subject variances, different within-subject variances between two methods and poor correlation between measurements of two methods.

Roy (2009) proposes the use of LME models to perform a test on two methods of agreement to determine whether they can be used interchangeably.

Roy (2009) proposes a LME based approach with Kronecker product covariance structure with doubly multivariate setup to assess the agreement between two methods. This method is designed such that the data may be unbalanced and with unequal numbers of replications for each subject.

Roy (2009) uses an approach based on linear mixed effects (LME) models for the purpose of comparing the agreement between two methods of measurement, where replicate measurements on items (ofttimes individuals) by both methods are available. Three tests of hypothesis are provided, appropriate for evaluating the agreement between the two methods of measurement under this sampling scheme. These tests consider null hypotheses that assume: absence of inter-method bias; equality of between-subject variabilities of the two methods; equality of within-subject variabilities of the two methods. By inter-method bias we mean that a systematic difference exists between observations recorded by the two methods.

Roy proposes a novel method using the LME model with Kronecker product covariance structure in a doubly multivariate set-up to assess the agreement between a new method and an established method with unbalanced data and with unequal replications for different subjects (Roy, 2009).

Using Roy's method, four candidate models are constructed, each differing by constraints applied to the variance covariance matrices. In addition to computing the inter-method bias, three significance tests are carried out on the respective formulations to make a judgement on whether or not two methods are in agreement. Differences in between-subject variabilities of the two methods arise when one method is yield-

ing average response levels for individuals than are more variable than the average response levels for the same sample of individuals taken by the other method. Differences in within-subject variabilities of the two methods arise when one method is yielding responses for an individual than are more variable than the responses for this same individual taken by the other method. The two methods of measurement can be considered to agree, and subsequently can be used interchangeably, if all three null hypotheses are true.

For the purposes of comparing two methods of measurement, Roy (2009) presents a methodology utilizing linear mixed effects model. This methodology provides for the formal testing of inter-method bias, between-subject variability and within-subject variability of two methods. This formulation contains a Kronecker product covariance structure in a doubly multivariate setup. By doubly multivariate set up, Roy means that the information on each patient or item is multivariate at two levels, the number of methods and number of replicated measurements. Further to Lam et al. (1999), it is assumed that the replicates are linked over time. However it is easy to modify to the unlinked case.

Roy (2009) states three criteria for two methods to be considered in agreement. Firstly that there be no significant bias. Second that there is no difference in the between-subject variabilities, and lastly that there is no significant difference in the within-subject variabilities. Roy further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal. Variability tests proposed by Roy (2009) affords the opportunity to expand upon Carstensen’s approach.

### **3 Agreement Criteria**

Roy (2009) proposes a suite of hypothesis tests for assessing the agreement of two methods of measurement, when replicate measurements are obtained for each item,

using a LME approach. (An item would commonly be a patient).

Two methods of measurement can be said to be in agreement if there is no significant difference between in three key respects.

Firstly, there is no inter-method bias between the two methods, i.e. there is no persistent tendency for one method to give higher values than the other.

Secondly, both methods of measurement have the same within-subject variability. In such a case the variance of the replicate measurements would consistent for both methods. Lastly, the methods have equal between-subject variability. Put simply, for the mean measurements for each case, the variances of the mean measurements from both methods are equal.

Lack of agreement can arise if there is a disagreement in overall variabilities. This may be due to due to the disagreement in either between-item variabilities or within-item variabilities, or both. Roy (2009) allows for a formal test of each.

Roy (2009) sets out three criteria for two methods to be considered in agreement. Firstly that there be no significant bias. Second that there is no difference in the between-subject variabilities, and lastly that there is no significant difference in the within-subject variabilities. Roy further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal.

Two methods of measurement are in complete agreement if the null hypotheses  $H_1: \alpha_1 = \alpha_2$  and  $H_2: \sigma_1^2 = \sigma_2^2$  and  $H_3: g_1^2 = g_2^2$  hold simultaneously. Roy (2009) uses a Bonferroni correction to control the familywise error rate for tests of  $\{H_1, H_2, H_3\}$  and account for difficulties arising due to multiple testing. Roy also integrates  $H_2$  and  $H_3$  into a single testable hypothesis  $H_4: \omega_1^2 = \omega_2^2$ , where  $\omega_m^2 = \sigma_m^2 + g_m^2$  represent the overall variability of method  $m$ . Disagreement in overall variability may be caused by different between-item variabilities, by different within-item variabilities, or by both. If the exact cause of disagreement between the two methods is not of interest, then the overall variability test  $H_4$  is an alternative to testing  $H_2$  and  $H_3$  separately.



(Work this in) Roy’s method considers two methods to be in agreement if three conditions are met.

- no significant bias, i.e. the difference between the two mean readings is not ”statistically significant”,
- high overall correlation coefficient,
- the agreement between the two methods by testing their repeatability coefficients.

### 3.0.1 Inter-Method Bias

A formal test for inter-method bias can be implemented by examining the fixed effects of the model. This is common to well known classical linear model methodologies. The null hypotheses, that both methods have the same mean, which is tested against the alternative hypothesis, that both methods have different means. The inter-method bias and necessary  $t$ -value and  $p$ -value are presented in computer output. A decision on whether the first of Roy’s criteria is fulfilled can be based on these values.

Importantly Roy (2009) further proposes a series of three tests on the variance components of an LME model, which allow decisions on the second and third of Barnhart’s criteria. For these tests, four candidate LME models are constructed. The differences in the models are specifically in how the  $D$  and  $\Lambda$  matrices are constructed, using either an unstructured form or a compound symmetry form. To illustrate these differences, consider a generic matrix  $A$ ,

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

A symmetric matrix allows the diagonal terms  $a_{11}$  and  $a_{22}$  to differ. The compound symmetry structure requires that both of these terms be equal, i.e  $a_{11} = a_{22}$ . The first model acts as an alternative hypothesis to be compared against each of three other models, acting as null hypothesis models, successively. The models are compared using the likelihood ratio test. Likelihood ratio tests are a class of tests based on the

comparison of the values of the likelihood functions of two candidate models. LRTs can be used to test hypotheses about covariance parameters or fixed effects parameters in the context of LMEs. The test statistic for the likelihood ratio test is the difference of the log-likelihood functions, multiplied by  $-2$ . The probability distribution of the test statistic is approximated by the  $\chi^2$  distribution with  $(\nu_1 - \nu_2)$  degrees of freedom, where  $\nu_1$  and  $\nu_2$  are the degrees of freedom of models 1 and 2 respectively. Each of these three test shall be examined in due course.

## 4 Roy's Hypotheses Tests

In order to express Roy's LME model in matrix notation we gather all  $2n_i$  observations specific to item  $i$  into a single vector  $\mathbf{y}_i = (y_{1i1}, y_{2i1}, y_{1i2}, \dots, y_{mir}, \dots, y_{1in_i}, y_{2in_i})'$ . The LME model can be written

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i,$$

where  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$  is a vector of fixed effects, and  $\mathbf{X}_i$  is a corresponding  $2n_i \times 3$  design matrix for the fixed effects. The random effects are expressed in the vector  $\mathbf{b} = (b_1, b_2)'$ , with  $\mathbf{Z}_i$  the corresponding  $2n_i \times 2$  design matrix. The vector  $\boldsymbol{\epsilon}_i$  is a  $2n_i \times 1$  vector of residual terms.

It is assumed that  $\mathbf{b}_i \sim N(0, \mathbf{G})$ ,  $\boldsymbol{\epsilon}_i$  is a matrix of random errors distributed as  $N(0, \mathbf{R}_i)$  and that the random effects and residuals are independent of each other.

$\mathbf{G}$  is the variance covariance matrix for the random effects  $\mathbf{b}$ . i.e. between-item sources of variation. The between-item variance covariance matrix  $\mathbf{G}$  is constructed as follows:

The distribution of the random effects is described as  $\mathbf{b}_i \sim N(0, \mathbf{G})$ . Similarly random errors are distributed as  $\boldsymbol{\epsilon}_i \sim N(0, \mathbf{R}_i)$ . The random effects and residuals are assumed to be independent. Both covariance matrices can be written as follows;

$$\mathbf{G} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix}$$

and

$$\mathbf{R}_i = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & 0 & 0 & \dots & \dots & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & 0 & 0 & \dots & \dots & 0 & 0 \\ 0 & 0 & \sigma_1^2 & \sigma_{12} & \dots & \dots & 0 & 0 \\ 0 & 0 & \sigma_{12} & \sigma_2^2 & \dots & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \dots & \sigma_1^2 & \sigma_{12} \\ 0 & 0 & 0 & 0 & \dots & \dots & \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

$$\text{Var} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \mathbf{G} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix}$$

It is important to note that no special assumptions about the structure of  $\mathbf{G}$  are made. An example of such an assumption would be that  $\mathbf{G}$  is the product of a scalar value and the identity matrix.

$\mathbf{R}_i$  is the variance covariance matrix for the residuals, i.e. the within-item sources of variation between both methods. Computational analysis of linear mixed effects models allow for the explicit analysis of both  $\mathbf{G}$  and  $\mathbf{R}_i$ . The above terms can be used to express the variance covariance matrix  $\mathbf{\Omega}_i$  for the responses on item  $i$ ,

$$\mathbf{\Omega}_i = \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i' + \mathbf{R}_i.$$

Hamlett et al. (2004) shows that  $\mathbf{R}_i$  can be expressed as  $\mathbf{R}_i = \mathbf{I}_{n_i} \otimes \mathbf{\Sigma}$ . The partial within-item variance?covariance matrix of two methods at any replicate is denoted  $\mathbf{\Sigma}$ , where  $\sigma_1^2$  and  $\sigma_2^2$  are the within-subject variances of the respective methods, and  $\sigma_{12}$  is the within-item covariance between the two methods. It is assumed that the within-item variance?covariance matrix  $\mathbf{\Sigma}$  is the same for all replications. Again it is important to note that no special assumptions are made about the structure of the matrix.

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \quad (3)$$

For expository purposes consider the case where each item provides three replicates by each method. Then in matrix notation the model has the structure

$$\mathbf{y}_i = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} + \begin{pmatrix} \epsilon_{1i1} \\ \epsilon_{2i1} \\ \epsilon_{1i2} \\ \epsilon_{2i2} \\ \epsilon_{1i3} \\ \epsilon_{2i3} \end{pmatrix}, \quad (4)$$

where

$$\mathbf{G} =$$

and

$$\mathbf{R}_i =$$

It is assumed that  $\mathbf{b}_i \sim N(0, \mathbf{G})$ ,  $\boldsymbol{\epsilon}_i$  is a matrix of random errors distributed as  $N(0, \mathbf{R}_i)$  and that the random effects and residuals are independent of each other. Assumptions made on the structures of  $\mathbf{G}$  and  $\mathbf{R}_i$  will be discussed in due course.

The partial within-item variance covariance matrix of two methods at any replicate is denoted  $\boldsymbol{\Sigma}$ , where  $\sigma_1^2$  and  $\sigma_2^2$  are the within-subject variances of both methods, and  $\sigma_{12}$  is the within-item covariance between the two methods. The within-item variance covariance matrix  $\boldsymbol{\Sigma}$  is assumed to be the same for all replications.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

The overall variability between the two methods is the sum of between-item variability  $\mathbf{G}$  and within-item variability  $\boldsymbol{\Sigma}$ . Roy (2009) denotes the overall variability as Block -  $\boldsymbol{\Omega}_i$ . The overall variation for methods 1 and 2 are given by

$$\begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

Hamlett et al. (2004) shows that  $\mathbf{R}_i$  can be expressed as  $\mathbf{I}_{n_i} \otimes \mathbf{\Sigma}$ . The covariance matrix has the same structure for all items, except for dimension, which depends on the number of replicates. The  $2 \times 2$  block diagonal Block- $\mathbf{\Omega}_i$  represents the covariance matrix between two methods, and is the sum of  $\mathbf{G}$  and  $\mathbf{\Sigma}$ .

$$\text{Block-}\mathbf{\Omega}_i = \begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

The variance of case-wise difference in measurements can be determined from Block- $\mathbf{\Omega}_i$ . Hence limits of agreement can be computed.

The computation of the limits of agreement require that the variance of the difference of measurements. This variance is easily computable from the estimate of the Block -  $\mathbf{\Omega}_i$  matrix. Lack of agreement can arise if there is a disagreement in overall variabilities. This may be due to the disagreement in either between-item variabilities or within-item variabilities, or both. Roy (2009) allows for a formal test of each.

## 4.1 Model Specification

Let  $y_{mir}$  be the  $r$ th replicate measurement on the  $i$ th item by the  $m$ th method, where  $m = 1, 2$ ,  $i = 1, \dots, N$ , and  $r = 1, \dots, n_i$ . When the design is balanced and there is no ambiguity we can set  $n_i = n$ . The LME model can be written

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \quad (5)$$

Here  $\beta_0$  and  $\beta_m$  are fixed-effect terms representing, respectively, a model intercept and an overall effect for method  $m$ . The  $b_{1i}$  and  $b_{2i}$  terms represent random effect parameters corresponding to the two methods, having  $E(b_{mi}) = 0$  with  $\text{Var}(b_{mi}) = g_m^2$  and  $\text{Cov}(b_{mi}, b_{m'i}) = g_{12}$ . The random error term for each response is denoted  $\epsilon_{mir}$  having  $E(\epsilon_{mir}) = 0$ ,  $\text{Var}(\epsilon_{mir}) = \sigma_m^2$ ,  $\text{Cov}(b_{mir}, b_{m'ir}) = \sigma_{12}$ ,  $\text{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$  and  $\text{Cov}(\epsilon_{mir}, \epsilon_{m'ir'}) = 0$ . When two methods of measurement are in agreement, there is no significant differences between  $\beta_1$  and  $\beta_2$ ,  $g_1^2$  and  $g_2^2$ , and  $\sigma_1^2$  and  $\sigma_2^2$ . Here  $\beta_0$  and  $\beta_m$  are fixed-effect terms representing, respectively, a model intercept and an overall effect for method  $m$ . The model can be reparameterized by gathering the  $\beta$  terms together into (fixed effect) intercept terms  $\alpha_m = \beta_0 + \beta_m$ . The  $b_{1i}$  and  $b_{2i}$  terms are correlated random effect parameters having  $E(b_{mi}) = 0$  with  $\text{Var}(b_{mi}) = g_m^2$  and  $\text{Cov}(b_{1i}, b_{2i}) = g_{12}$ . The random error term for each response is denoted  $\epsilon_{mir}$  having  $E(\epsilon_{mir}) = 0$ ,  $\text{Var}(\epsilon_{mir}) = \sigma_m^2$ ,  $\text{Cov}(\epsilon_{1ir}, \epsilon_{2ir}) = \sigma_{12}$ ,  $\text{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$  and  $\text{Cov}(\epsilon_{1ir}, \epsilon_{2ir'}) = 0$ .

### 4.1.1 Model Terms (Roy 2009)

It is important to note the following characteristics of this model.

- Let the number of replicate measurements on each item  $i$  for both methods be  $n_i$ , hence  $2 \times n_i$  responses. However, it is assumed that there may be a different number of replicates made for different items. Let the maximum number of replicates be  $p$ . An item will have up to  $2p$  measurements, i.e.  $\max(n_i) = 2p$ .
- Later on  $\mathbf{X}_i$  will be reduced to a  $2 \times 1$  matrix, to allow estimation of terms. This is due to a shortage of rank. The fixed effects vector can be modified accordingly.

- $\mathbf{Z}_i$  is the  $2n_i \times 2$  model matrix for the random effects for measurement methods on item  $i$ .
- $\mathbf{b}_i$  is the  $2 \times 1$  vector of random-effect coefficients on item  $i$ , one for each method.
- $\boldsymbol{\epsilon}$  is the  $2n_i \times 1$  vector of residuals for measurements on item  $i$ .
- $\mathbf{G}$  is the  $2 \times 2$  covariance matrix for the random effects.
- $\mathbf{R}_i$  is the  $2n_i \times 2n_i$  covariance matrix for the residuals on item  $i$ .
- The expected value is given as  $E(\mathbf{y}_i) = \mathbf{X}_i\boldsymbol{\beta}$ . (Hamlett et al., 2004)
- The variance of the response vector is given by  $\text{Var}(\mathbf{y}_i) = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$  (Hamlett et al., 2004).

The maximum likelihood estimate of the between-subject variance covariance matrix of two methods is given as  $D$ . The estimate for the within-subject variance covariance matrix is  $\hat{\Sigma}$ . The estimated overall variance covariance matrix ‘Block  $\Omega_i$ ’ is the addition of  $\hat{D}$  and  $\hat{\Sigma}$ .

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \quad (6)$$

- $\mathbf{b}_i$  is a  $m$ -dimensional vector comprised of the random effects.

$$\mathbf{b}_i = \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} \quad (7)$$

- $\mathbf{V}$  represents the correlation matrix of the replicated measurements on a given method.  $\boldsymbol{\Sigma}$  is the within-subject VC matrix.
- $\mathbf{V}$  and  $\boldsymbol{\Sigma}$  are positive definite matrices. The dimensions of  $\mathbf{V}$  and  $\boldsymbol{\Sigma}$  are  $3 \times 3 (= p \times p)$  and  $2 \times 2 (= k \times k)$ .
- It is assumed that  $\mathbf{V}$  is the same for both methods and  $\boldsymbol{\Sigma}$  is the same for all replications.
- $\mathbf{V} \otimes \boldsymbol{\Sigma}$  creates a  $6 \times 6 (= kp \times kp)$  matrix.  $\mathbf{R}_i$  is a sub-matrix of this.



## 4.2 Differences Between Approaches : Assumptions on Variability

Aside from the fixed effects, another important difference is that Carstensen’s model requires that particular assumptions be applied, specifically that the off-diagonal elements of the between-item and within-item variability matrices are zero. By extension the overall variability off diagonal elements are also zero.

Also, implementation requires that the between-item variances are estimated as the same value:  $g_1^2 = g_2^2 = g^2$ . Necessarily Carstensen’s method does not allow for a formal test of the between-item variability.

$$\begin{pmatrix} \omega_2^1 & 0 \\ 0 & \omega_2^2 \end{pmatrix} = \begin{pmatrix} g^2 & 0 \\ 0 & g^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

In cases where the off-diagonal terms in the overall variability matrix are close to zero, the limits of agreement due to ? are very similar to the limits of agreement that follow from the general model.

Carstensen et al. (2008) develop their model from a standard two-way analysis of variance model, reformulated for the case of replicate measurements, with random effects terms specified as appropriate. Their model describing  $y_{mir}$ , again the  $r$ th replicate measurement on the  $i$ th item by the  $m$ th method ( $m = 1, 2, i = 1, \dots, N$ , and  $r = 1, \dots, n$ ), can be written as

$$y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + \epsilon_{mir}. \quad (8)$$

The fixed effects  $\alpha_m$  and  $\mu_i$  represent the intercept for method  $m$  and the ‘true value’ for item  $i$  respectively. The random-effect terms comprise an item-by-replicate interaction term  $a_{ir} \sim \mathcal{N}(0, \varsigma^2)$ , a method-by-item interaction term  $c_{mi} \sim \mathcal{N}(0, \tau_m^2)$ , and model error terms  $\epsilon \sim \mathcal{N}(0, \varphi_m^2)$ . All random-effect terms are assumed to be independent. For the case when replicate measurements are assumed to be exchangeable for item  $i$ ,  $a_{ir}$  can be removed.

There is a substantial difference in the number of fixed parameters used by the respective models. For the model in (??) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items  $N$ . In contrast, the model described by (8) requires  $N + 2$  fixed effects for  $N$  items. The inclusion of fixed effects to account for the ‘true value’ of each item greatly increases the level of model complexity.

When only two methods are compared, Carstensen et al. (2008) notes that separate estimates of  $\tau_m^2$  can not be obtained due to the model over-specification. To overcome this, the assumption of equality, i.e.  $\tau_1^2 = \tau_2^2$ , is required.

### 4.3 Limits of Agreement in LME models

Carstensen et al. (2008) uses LME models to determine the limits of agreement. Between-subject variation for method  $m$  is given by  $d_m^2$  and within-subject variation is given by  $\lambda_m^2$ . Carstensen et al. (2008) remarks that for two methods  $A$  and  $B$ , separate values of  $d_A^2$  and  $d_B^2$  cannot be estimated, only their average. Hence the assumption that  $d_x = d_y = d$  is necessary. The between-subject variability  $\mathbf{D}$  and within-subject variability  $\mathbf{\Lambda}$  can be presented in matrix form,

$$\mathbf{D} = \begin{pmatrix} d_A^2 & 0 \\ 0 & d_B^2 \end{pmatrix} = \begin{pmatrix} d^2 & 0 \\ 0 & d^2 \end{pmatrix}, \quad \mathbf{\Lambda} = \begin{pmatrix} \lambda_A^2 & 0 \\ 0 & \lambda_B^2 \end{pmatrix}.$$

The variance for method  $m$  is  $d_m^2 + \lambda_m^2$ . Limits of agreement are determined using the standard deviation of the case-wise differences between the sets of measurements by two methods  $A$  and  $B$ , given by

$$\text{var}(y_A - y_B) = 2d^2 + \lambda_A^2 + \lambda_B^2. \quad (9)$$

Importantly the covariance terms in both variability matrices are zero, and no covariance component is present.

Roy (2009) has demonstrated a methodology whereby  $d_A^2$  and  $d_B^2$  can be estimated separately. Also covariance terms are present in both  $\mathbf{D}$  and  $\mathbf{\Lambda}$ . Using Roy's methodology, the variance of the differences is

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_B^2 + \lambda_A^2 - 2(d_{AB} + \lambda_{AB}) \quad (10)$$

All of these terms are given or determinable in computer output. The limits of agreement can therefore be evaluated using

$$\bar{y}_A - \bar{y}_B \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \quad (11)$$

## 4.4 Remarks

# 5 Roy's Hypothesis Testing

Variability tests proposed by Roy (2009) affords the opportunity to expand upon Carstensen's approach.

The first test allows of the comparison the begin-subject variability of two methods. Similarly, the second test assesses the within-subject variability of two methods. A third test is a test that compares the overall variability of the two methods.

The tests are implemented by fitting a specific LME model, and three variations thereof, to the data. These three variant models introduce equality constraints that act null hypothesis cases.

Other important aspects of the method comparison study are consequent. The limits of agreement are computed using the results of the first model.

The formulation presented above usefully facilitates a series of significance tests that advise as to how well the two methods agree. These tests are as follows:

- A formal test for the equality of between-item variances,
- A formal test for the equality of within-item variances,
- A formal test for the equality of overall variances.

These tests are complemented by the ability to consider the inter-method bias and the overall correlation coefficient. Two methods can be considered to be in agreement if criteria based upon these methodologies are met. Additionally Roy makes reference to the overall correlation coefficient of the two methods, which is determinable from variance estimates.

Lack of agreement can arise if there is a disagreement in overall variabilities. This lack of agreement may be due to differing between-item variabilities, differing within-item variabilities, or both. The formulation presented above usefully facilitates a series of significance tests that assess if and where such differences arise. Roy (2009) allows

for a formal test of each. These tests are comprised of a formal test for the equality of between-item variances,

$$H_0 : g_1^2 = g_2^2$$

$$H_1 : g_1^2 \neq g_2^2$$

a formal test for the equality of within-item variances,

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

and finally, a formal test for the equality of overall variances.

$$H_0 : \omega_1^2 = \omega_2^2$$

$$H_1 : \omega_1^2 \neq \omega_2^2$$

These tests are complemented by the ability to consider the inter-method bias and the overall correlation coefficient. Two methods can be considered to be in agreement if criteria based upon these methodologies are met. Additionally Roy makes reference to the overall correlation coefficient of the two methods, which is determinable from variance estimates.

## 6 Computation of limits of agreement under Roy's model

The limits of agreement (Bland and Altman, 1986) are ubiquitous in method comparison studies. The computation thereof require that the variance of the difference of measurements. This variance is easily computable from the variance estimates in the Block -  $\mathbf{\Omega}_i$  matrix, i.e.

$$\text{Var}(y_1 - y_2) = \sqrt{\omega_1^2 + \omega_2^2 - 2\omega_{12}}.$$

## 7 Carstensen's Limits of agreement

? presents a methodology to compute the limits of agreement based on LME models. Importantly, Carstensen's underlying model differs from Roy's model in some key respects, and therefore a prior discussion of Carstensen's model is required.

## 8 Roy's LME methodology for assessing agreement

Bivariate correlation coefficients have been shown to be of limited use in method comparison studies (Bland and Altman, 1986). However, recently correlation analysis has been developed to cope with repeated measurements, enhancing their potential usefulness. Roy incorporates the use of correlation into his methodology.

Roy (2009) considers the problem of assessing the agreement between two methods with replicate observations in a doubly multivariate set-up using linear mixed effects models.

Roy (2009) uses examples from Bland and Altman (1986) to be able to compare both types of analysis.

Roy (2009) considers four independent hypothesis tests.

- Testing of hypotheses of differences between the means of two methods
- Testing of hypotheses in between subject variabilities in two methods,
- Testing of hypotheses of differences in within-subject variability of the two methods,
- Testing of hypotheses in differences in overall variability of the two methods.

## 8.1 Roy's variability tests

The first test allows of the comparison the begin-subject variability of two methods. Similarly, the second test assesses the within-subject variability of two methods. A third test is a test that compares the overall variability of the two methods.

The tests are implemented by fitting a specific LME model, and three variations thereof, to the data. These three variant models introduce equality constraints that act null hypothesis cases.

Other important aspects of the method comparison study are consequent. The limits of agreement are computed using the results of the first model.

## 8.2 Hypothesis Testing

The formulation presented above usefully facilitates a series of significance tests that advise as to how well the two methods agree. These tests are as follows:

- A formal test for the equality of between-item variances,
- A formal test for the equality of within-item variances,
- A formal test for the equality of overall variances.

These tests are complemented by the ability to consider the inter-method bias and the overall correlation coefficient. Two methods can be considered to be in agreement if criteria based upon these methodologies are met. Additionally Roy makes reference to the overall correlation coefficient of the two methods, which is determinable from variance estimates.

Variability tests proposed by Roy (2009) affords the opportunity to expand upon Carstensen's approach.

The first test allows of the comparison the begin-subject variability of two methods. Similarly, the second test assesses the within-subject variability of two methods. A third test is a test that compares the overall variability of the two methods.



The tests are implemented by fitting a specific LME model, and three variations thereof, to the data. These three variant models introduce equality constraints that act null hypothesis cases.

Other important aspects of the method comparison study are consequent. The limits of agreement are computed using the results of the first model.

### 8.3 Variability test 1

The first test determines whether or not both methods  $A$  and  $B$  have the same between-subject variability, further to the second of Roy's criteria.

$$H_0 : d_A = d_B$$

$$H_A : d_A \neq d_B$$

This test is facilitated by constructing a model specifying a symmetric form for  $D$  (i.e. the alternative model) and comparing it with a model that has compound symmetric form for  $D$  (i.e. the null model). For this test  $\hat{\Lambda}$  has a symmetric form for both models, and will be the same for both.

### 8.4 Variability test 2

This test determines whether or not both methods  $A$  and  $B$  have the same within-subject variability, thus enabling a decision on the third of Roy's criteria.

$$H_0 : \lambda_A = \lambda_B$$

$$H_A : \lambda_A \neq \lambda_B$$

### 8.5 Correlation

Bivariate correlation coefficients have been shown to be of limited use in method comparison studies (Bland and Altman, 1986). However, recently correlation analysis has

been developed to cope with repeated measurements, enhancing their potential usefulness. Roy incorporates the use of correlation into his methodology.

In addition to the variability tests, Roy advises that it is preferable that a correlation of greater than 0.82 exist for two methods to be considered interchangeable. However if two methods fulfil all the other conditions for agreement, failure to comply with this one can be overlooked. Indeed Roy demonstrates that placing undue importance to it can lead to incorrect conclusions. Roy (2009) remarks that current computer implementations only gives overall correlation coefficients, but not their variances. Consequently it is not possible to carry out inferences based on all overall correlation coefficients.

- Let  $y_{mir}$  be the response of method  $m$  on the  $i$ th subject at the  $r$ —th replicate.
- Let  $\mathbf{y}_{ir}$  be the  $2 \times 1$  vector of measurements corresponding to the  $i$ —th subject at the  $r$ —th replicate.
- Let  $\mathbf{y}_i$  be the  $R_i \times 1$  vector of measurements corresponding to the  $i$ —th subject, where  $R_i$  is number of replicate measurements taken on item  $i$ .
- Let  $\alpha_{mi}$  be the fixed effect parameter for method for subject  $i$ .
- Formally Roy uses a separate fixed effect parameter to describe the true value  $\mu_i$ , but later combines it with the other fixed effects when implementing the model.
- Let  $u_{1i}$  and  $u_{2i}$  be the random effects corresponding to methods for item  $i$ .
- $\boldsymbol{\epsilon}_i$  is a  $n_i$ -dimensional vector comprised of residual components. For the blood pressure data  $n_i = 85$ .
- $\boldsymbol{\beta}$  is the solutions of the means of the two methods. In the LME output, the bias and corresponding t-value and p-values are presented. This is relevant to Roy's first test.

## 9 Roy's LME approach

The methodology uses a linear mixed effects regression fit using compound symmetry (CS) correlation structure on  $\mathbf{V}$ .

$$\Lambda = \frac{\max_{H_0} L}{\max_{H_1} L}$$

Roy (2009) considers the problem of assessing the agreement between two methods with replicate observations in a doubly multivariate set-up using linear mixed effects models.

Roy (2009) uses examples from Bland and Altman (1986) to be able to compare both types of analysis.

## 10 Roy’s LME methodology for assessing agreement

Barnhart et al. (2007) describes the sources of disagreement as differing population means, different between-subject variances, different within-subject variances between two methods and poor correlation between measurements of two methods.

Roy (2009) proposes the use of LME models to perform a test on two methods of agreement to determine whether they can be used interchangeably. Roy (2009) considers the problem of assessing the agreement between two methods with replicate observations in a doubly multivariate set-up using linear mixed effects models.

Roy (2009) uses examples from Bland and Altman (1986) to be able to compare both types of analysis.

Roy (2009) proposes a LME based approach with Kronecker product covariance structure with doubly multivariate setup to assess the agreement between two methods. This method is designed such that the data may be unbalanced and with unequal numbers of replications for each subject.

Roy (2009) considers four independent hypothesis tests.

- Testing of hypotheses of differences between the means of two methods
- Testing of hypotheses in between subject variabilities in two methods,
- Testing of hypotheses of differences in within-subject variability of the two methods,
- Testing of hypotheses in differences in overall variability of the two methods.

### 10.0.1 Sampling Scheme : Replicates

Measurements taken in quick succession by the same observer using the same instrument on the same subject can be considered true replicates. Roy (2009) notes that some measurements may not be ‘true’ replicates.

Roy’s methodology assumes the use of ‘true replicates’. However data may not be collected in this way. In such cases, the correlation matrix on the replicates may require

a different structure, such as the autoregressive order one  $AR(1)$  structure. However determining MLEs with such a structure would be computational intense, if possible at all.

## 10.1 Difference Variance further to Carstensen

Carstensen et al. (2008) states a model where the variation between items for method  $m$  is captured by  $\tau_m$  (our notation  $d_m^2$ ) and the within-item variation by  $\sigma_m$ .

*The formulation of this model is general and refers to comparison of any number of methods however, if only two methods are compared, separate values of  $\tau_1^2$  and  $\tau_2^2$  cannot be estimated, only their average value  $\tau$ , so in the case of only two methods we are forced to assume that  $\tau_1 = \tau_2 = \tau(?)$ .*

Another important point is that there is no covariance terms, so further to ? the variance covariance matrices for between-item and within-item variability are respectively.

$$\mathbf{D} = \begin{pmatrix} d_2^1 & 0 \\ 0 & d_2^2 \end{pmatrix}$$

and  $\mathbf{\Sigma}$  is constructed as follows:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_2^1 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

Under this model the limits of agreement should be computed based on the standard deviation of the difference between a pair of measurements by the two methods on a new individual,  $j$ , say:

$$\text{var}(y_{1j} - y_{2j}) = 2d^2 + \sigma_1^2 + \sigma_2^2$$

Further to his model, Carstensen computes the limits of agreement as

$$\hat{\alpha}_1 - \hat{\alpha}_2 \pm \sqrt{2\hat{d}^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2}$$

## 10.2 Relevance of Roy’s Methodology

The relevance of Roy’s methodology is that estimates for the between-item variances for both methods  $\hat{d}_m^2$  are computed. Also the VC matrices are constructed with covariance terms and, so the difference variance must be formulated accordingly.

$$\hat{\alpha}_1 - \hat{\alpha}_2 \pm \sqrt{\hat{d}_1^2 + \hat{d}_2^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{d}_{12} - 2\hat{\sigma}_{12}}$$

## References

- Barnhart, H., M. Haber, and L. Lin (2007). An overview of assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics* 17, 529–569.
- Bland, J. and D. Altman (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* i, 307–310.
- Bland, J. and D. Altman (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 8(2), 135–160.
- Carstensen, B. (2004). Comparing and predicting between several methods of measurement. *Biostatistics* 5(3), 399–413.
- Carstensen, B., J. Simpson, and L. C. Gurrin (2008). Statistical models for assessing agreement in method comparison studies with replicate measurements. *The International Journal of Biostatistics* 4(1).
- Hamlett, A., L. Ryan, and R. Wolfinger (2004). On the use of PROC MIXED to estimate correlation in the presence of repeated measures. *Proceedings of the Statistics and Data Analysis Section, SAS Users Group International* 198-229, 1–7.
- Lai, D. and S.-Y. P. K. Shiao (2005). Comparing two clinical measurements: a linear mixed model approach. *Journal of Applied Statistics* 32(8), 855–860.

- Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics* 38(4), 963–974.
- Lam, M., K. Webb, and D. O'Donnell (1999). Correlation between two variables in repeated measurements. *American Statistical Association, Proceedings of the Biometric Session*, 213–218.
- Roy, A. (2009). An application of the linear mixed effects model to ass the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics* 19, 150–173.