

# Contents

<b>1</b>	<b>LOAS</b>	<b>3</b>
1.1	Limits of agreement in LME models . . . . .	3
1.2	Computing LoAs from LME models . . . . .	8
1.3	Carstensen's Mixed Models . . . . .	8
1.4	Carstensen's Model . . . . .	10
1.5	Calculation of limits of agreement . . . . .	12
1.6	Extension of Roy's methodology . . . . .	13
1.7	Roy's methodology for single measurements . . . . .	14
1.8	Correlation . . . . .	15
1.9	Correlation terms . . . . .	15
1.10	Limits of agreement in LME models . . . . .	16
1.10.1	Variance Ratios . . . . .	17
1.11	Testing Procedures . . . . .	18
1.12	Carstensen's Limits of Agreement . . . . .	18
1.13	BXC2008 presents - Carstensen's Limits of agreement . . . . .	19
1.14	Carstensen's LOAs . . . . .	20
1.15	Interaction Terms in Model . . . . .	20
1.16	Computation of limits of agreement . . . . .	21
1.17	Computation of limits of agreement under Roy's model . . . . .	22
1.18	Difference Between Approaches . . . . .	23
1.19	Differences Between Models . . . . .	24

1.20	Differences . . . . .	24
1.21	Relevance of Roy's Methodology . . . . .	25
1.22	Difference Variance further to Carstensen . . . . .	26
1.23	Assumptions on Variability . . . . .	27

# Chapter 1

## LOAS

### 1.1 Limits of agreement in LME models

Limits of agreement are used extensively for assessing agreement, because they are intuitive and easy to use. Necessarily their prevalence in literature has meant that they are now the best known measurement for agreement, and therefore any newer methodology would benefit by making reference to them.

In computing limits of agreement, it is first necessary to have an estimate for the standard deviations of the differences. When the agreement of two methods is analyzed using LME models, a clear method of how to compute the standard deviation is required. As the estimate for inter-method bias and the quantile would be the same for both methodologies, the focus hereon is solely on the variance of differences.

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, they recommend 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 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

$$\text{var}(y_A - y_B) = 2d^2 + \lambda_A^2 + \lambda_B^2. \quad (1.1)$$

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

Roy (2009b) 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

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_A^2 + \lambda_B^2 - 2(d_{AB} + \lambda_{AB}) \quad (1.2)$$

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

$$\bar{y}_A - \bar{y}_B \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \quad (1.3)$$

The limits of agreement computed by Roy’s method are derived from the variance covariance matrix for overall variability. This matrix is the sum of the between subject VC matrix and the within-subject VC matrix.

The standard deviation of the differences of methods  $x$  and  $y$  is computed using values from the overall VC matrix.

$$\text{var}(x - y) = \text{var}(x) + \text{var}(y) - 2\text{cov}(x, y)$$

The respective estimates computed by both methods are tabulated as follows. Evidently there is close correspondence between both sets of estimates.

The computation thereof require that the variance of the difference of measurements. This variance is easily computable from the variance estimates in the Block -  $\Omega_i$  matrix, i.e.

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

Limits of agreement are used extensively for assessing agreement, because they are intuitive and easy to use. Necessaire their prevalence in literature has meant that they are now the best known measurement for agreement, and therefore any newer methodology would benefit by making reference to them.

However, the original Bland-Altman method was developed for two sets of measurements done on one occasion (i.e. independent data), and so this approach is not suitable for replicate measures data. However, as a naive analysis, it may be used to explore the data because of the simplicity of the method.

Bland and Altman (1999) addresses the issue of computing LoAs in the presence of replicate measurements, suggesting several computationally simple approaches. When repeated measures data are available, it is desirable to use all the data to compare the two methods. Further to Bland and Altman (1986), the computation of the limits of agreement follows from the intermethod bias, and the variance of the difference of measurements. The computation of the inter-method bias is a straightforward subtraction calculation. The variance of differences is easily computable from the variance

estimates in the Block -  $\Omega_i$  matrix, i.e.

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

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.

Carstensen et al. (2008) computes the limits of agreement to the case with replicate measurements by using LME models. As their interest mainly lies in extending the Bland-Altman methodology, other formal tests are not considered.

Carstensen et al. (2008) also presents a methodology to compute the limits of agreement based on LME models. In many cases the limits of agreement derived from this method accord with those to Roy's model. However, in other cases dissimilarities emerge. An explanation for this differences can be found by considering how the respective models account for covariance in the observations.

Specifying the relevant terms using a bivariate normal distribution, Roy's model allows for both between-method and within-method covariance. Carstensen et al. (2008) formulate a model whereby random effects have univariate normal distribution, and no allowance is made for correlation between observations.

A consequence of this is that the between-method and within-method covariance are zero. In cases where there is negligible covariance between methods, both sets of limits of agreement are very similar to each other. In cases where there is a substantial level of covariance present between the two methods, the limits of agreement computed using models will differ.

Carstensen et al. (2008) 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. The method of computation is the same as Roy's model, but with the

covariance estimates set to zero.

Carstensen et al. (2008) uses LME models to determine the limits of agreement. In computing limits of agreement, it is first necessary to have an estimate for the standard deviations of the differences. When the agreement of two methods is analyzed using LME models, a clear method of how to compute the standard deviation is required. As the estimate for inter-method bias and the quantile would be the same for both methodologies, the focus is solely on the standard deviation.

In cases where there is negligible covariance between methods, the limits of agreement computed using Roy’s model accord with those computed using Carstensen’s model. In cases where some degree of covariance is present between the two methods, the limits of agreement computed using models will differ. In the presented example, it is shown that Roy’s LoAs are lower than those of Carstensen, when covariance is present.

Importantly, estimates required to calculate the limits of agreement are not extractable, and therefore the calculation must be done by hand.

Carstensen presents a model where the variation between items for method  $m$  is captured by  $\sigma_m$  and the within item variation by  $\tau_m$ . Further to his model, Carstensen computes the limits of agreement as

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

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

$$\text{var}(y_A - y_B) = 2d^2 + \lambda_A^2 + \lambda_B^2. \quad (1.4)$$

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

Carstensen et al. (2008) presents a data set ‘fat’, which is a comparison of measurements of subcutaneous fat by two observers at the Steno Diabetes Center, Copenhagen.

Measurements are in millimeters (mm). Each person is measured three times by each observer. The observations are considered to be ‘true’ replicates.

A linear mixed effects model is formulated, and implementation through several software packages is demonstrated. All of the necessary terms are presented in the computer output. The limits of agreement are therefore,

$$0.0449 \pm 1.96 \times \sqrt{2 \times 0.0596^2 + 0.0772^2 + 0.0724^2} = (-0.220, 0.309). \quad (1.5)$$

Roy (2009a) 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

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_A^2 + \lambda_B^2 - 2(d_{AB} + \lambda_{AB}) \quad (1.6)$$

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

$$\bar{y}_A - \bar{y}_B \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \quad (1.7)$$

For Carstensen’s ‘fat’ data, the limits of agreement computed using Roy’s method are consistent with the estimates given by Carstensen et al. (2008);  $0.044884 \pm 1.96 \times 0.1373979 = (-0.224, 0.314)$ .

## 1.2 Computing LoAs from LME models

*One important feature of replicate observations is that they should be independent of each other. In essence, this is achieved by ensuring that the observer makes each measurement independent of knowledge of the previous value(s). This may be difficult to achieve in practice.*

## 1.3 Carstensen’s Mixed Models

Carstensen (2004) proposes linear mixed effects models for deriving conversion calculations similar to Deming’s regression, and for estimating variance components for



measurements by different methods. The following model ( in the authors own notation) is formulated as follows, where  $y_{mir}$  is the  $r$ th replicate measurement on subject  $i$  with method  $m$ .

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \quad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \quad (1.8)$$

The intercept term  $\alpha$  and the  $\beta_m \mu_i$  term follow from Dunn (2002), expressing constant and proportional bias respectively , in the presence of a real value  $\mu_i$ .  $c_{mi}$  is a interaction term to account for replicate, and  $e_{mir}$  is the residual associated with each observation. Since variances are specific to each method, this model can be fitted separately for each method.

The above formulation doesn't require the data set to be balanced. However, it does require a sufficient large number of replicates and measurements to overcome the problem of identifiability. The import of which is that more than two methods of measurement may be required to carry out the analysis. There is also the assumptions that observations of measurements by particular methods are exchangeable within subjects. (Exchangeability means that future samples from a population behaves like earlier samples).

Carstensen (2004) uses the above formula to predict observations for a specific individual  $i$  by method  $m$ ;

$$BLUP_{mir} = \hat{\alpha}_m + \hat{\beta}_m \mu_i + c_{mi} \quad (1.9)$$

. Under the assumption that the  $\mu$ s are the true item values, this would be sufficient to estimate parameters. When that assumption doesn't hold, regression techniques (known as updating techniques) can be used additionally to determine the estimates. The assumption of exchangeability can be unrealistic in certain situations. Carstensen (2004) provides an amended formulation which includes an extra interaction term ( $d_{mr} \sim N(0, \omega_m^2)$ )to account for this.

Carstensen et al. (2008) sets out a methodology of computing the limits of agreement based upon variance component estimates derived using linear mixed effects models. Measures of repeatability, a characteristic of individual methods of measurements, are also derived using this method.

Carstensen (2004) also advocates the use of linear mixed models in the study of method comparisons. The model is constructed 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;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \quad (e_{mi} \sim N(0, \sigma_m^2)) \quad (1.10)$$

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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \quad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \quad (1.11)$$

Carstensen et al. (2008) proposes a methodology to calculate prediction intervals in the presence of replicate measurements, overcoming problems associated with Bland-Altman methodology in this regard. It is not possible to estimate the interaction variance components  $\tau_1^2$  and  $\tau_2^2$  separately. Therefore it must be assumed that they are equal. The variance of the difference can be estimated as follows:

$$var(y_{1j} - y_{2j}) \quad (1.12)$$

## 1.4 Carstensen's Model

Using Carstensen's notation, a measurement  $y_{mi}$  by method  $m$  on individual  $i$  the measurement  $y_{mir}$  is the  $r$ th replicate measurement on the  $i$ th item by the  $m$ th method,

where  $m = 1, 2$ ,  $i = 1, \dots, N$ , and  $r = 1, \dots, n_i$  is formulated as follows;

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + \epsilon_{mir}, \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \quad (1.13)$$

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.

The classical model is based on measurements  $y_{mi}$  by method  $m = 1, 2$  on item  $i = 1, 2, \dots$

$$y_{mi} = \alpha_m + \mu_i + e_{mi}$$

$$e_{mi} \sim N(0, \sigma_m^2)$$

Here the terms  $\alpha_m$  and  $\mu_i$  represent the fixed effect for method  $m$  and a true value for item  $i$  respectively. The random effect terms comprise an interaction term  $c_{mi}$  and the residuals  $\epsilon_{mir}$ . The  $c_{mi}$  term represent random effect parameters corresponding to the two methods, having  $E(c_{mi}) = 0$  with  $\text{Var}(c_{mi}) = \tau_m^2$ . Carstensen specifies the variance of the interaction terms as being univariate normally distributed. As such,  $\text{Cov}(c_{mi}, c_{m'i}) = 0$ . All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method.

Even though the separate variances can not be identified, their sum can be estimated by the empirical variance of the differences.

Like wise the separate  $\alpha$  can not be estimated, only their difference can be estimated as  $\bar{D}$

The presence of the true value term  $\mu_i$  gives rise to an important difference between Carstensen's and Roy's models. 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. In other words, Roy considers the group of items being measured

as a sample taken from a population. Therefore 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.

With regards to specifying the variance terms, Carstensen remarks that using his approach is common, remarking that *The only slightly non-standard (meaning "not often used") feature is the differing residual variances between methods* (Carstensen, 2010).

Carstensen et al. (2008) makes some interesting remarks in this regard.

The only slightly non-standard (meaning "not often used") feature is the differing residual variances between methods.

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

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

As the difference between methods is of interest, the item term can be disregarded.

We assume that that the variance of the measurements is different for both methods, but it does not mean that the separate variances can be estimated with the data available.

## 1.5 Calculation of limits of agreement

The limits of agreement (Bland and Altman, 1986) are ubiquitous in method comparison studies.

However, the original BlandAltman method was developed for two sets of measurements done on one occasion (i.e. independent data), and so this approach is not suitable for replicate measures data. However, as a naive analysis, it may be used to explore the data because of the simplicity of the method.

The limits of agreement (Bland and Altman, 1986) are ubiquitous in method comparison studies.

Limits of agreement are used extensively for assessing agreement, because they are intuitive and easy to use.

Carstensen et al. (2008) computes the limits of agreement to the case with repeated measurements by using LME models.

Roy (2009b) formulates a very powerful method of assessing whether two methods of measurement, with replicate measurements, also using LME models. Roy's approach is based on the construction of variance-covariance matrices.

Importantly, Roy's approach does not address the issue of limits of agreement (though another related analysis, the coefficient of repeatability, is mentioned).

This paper seeks to use Roy's approach to estimate the limits of agreement. These estimates will be compared to estimates computed under Carstensen's formulation.

## 1.6 Extension of Roy's methodology

Roy's methodology is constructed to compare two methods in the presence of replicate measurements. Necessarily it is worth examining whether this methodology can be adapted for different circumstances.

An implementation of Roy's methodology, whereby three or more methods are used, is not feasible due to computational restrictions. Specifically there is a failure to reach convergence before the iteration limit is reached. This may be due to the presence of additional variables, causing the problem of non-identifiability. In the case of two variables, it is required to estimate two variance terms and four correlation terms, six in all. For the case of three variabilities, three variance terms must be estimated as well as nine correlation terms, twelve in all. In general for  $n$  methods has  $2 \times T_n$  variance terms, where  $T_n$  is the triangular number for  $n$ , i.e. the addition analogue of the factorial. Hence the computational complexity quite increases substantially for every increase in  $n$ .

Should an implementation be feasible, further difficulty arises when interpreting the results. The fundamental question is whether two methods have close agreement so as to be interchangeable. When three methods are present in the model, the null hypothesis is that all three methods have the same variability relevant to the respective tests. The outcome of the analysis will either be that all three are interchangeable or that all three are not interchangeable.

The tests would not be informative as to whether any two of those three were interchangeable, or equivalently if one method in particular disagreed with the other two. Indeed it is easier to perform three pair-wise comparisons separately and then to combine the results.

## 1.7 Roy's methodology for single measurements

Roy's methodology follows from the decomposition for the covariance matrix of the response vector  $y_i$ , as presented in Hamlett et al. (2004). The decomposition depends on the estimation of correlation terms, which would be absent in the single measurement case. Indeed there can be no within-subject variability if there are no repeated terms for it to describe. There would only be the covariance matrix of the measurements by both methods, which doesn't require the use of LME models. To conclude, simple existing methodologies would be the correct approach where there only one measurements by each method. Roy's methodology is not suitable for the case of single measurements because it follows from the decomposition for the covariance matrix of the response vector  $y_i$ , as presented in Hamlett et al. (2004). The decomposition depends on the estimation of correlation terms, which would be absent in the single measurement case. Indeed there can be no within-subject variability if there are no repeated terms for it to describe. There would only be the covariance matrix of the measurements by both methods, which doesn't require the use of LME models. To conclude, simpler existing methodologies, such as Deming regression, would be the correct approach where there only one measurements by each method.

Roy's methodology is not suitable for the case of single measurements because it follows from the decomposition for the covariance matrix of the response vector  $y_i$ , as presented in Hamlett et al. (2004). The decomposition depends on the estimation of correlation terms, which would be absent in the single measurement case. Indeed there can be no within-subject variability if there are no repeated terms for it to describe. There would only be the covariance matrix of the measurements by both methods, which doesn't require the use of LME models. To conclude, simpler existing methodologies, such as Deming regression, would be the correct approach where there only one measurements by each method.

## 1.8 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 (2009b) 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.

## 1.9 Correlation terms

The methodology proposed by ? is largely based on Hamlett et al. (2004), which in turn follows on from Lam et al. (1999).

Hamlett re-analyses the data of Lam et al. (1999) to generalize their model to cover other settings not covered by the Lam method.

In many cases, repeated observation are collected from each subject in sequence and/or longitudinally.

$$y_i = \alpha + \mu_i + \epsilon$$

Hamlett et al. (2004) demonstrated how the between-subject and within subject variabilities can be expressed in terms of correlation terms.

$$\mathbf{D} = \begin{pmatrix} \sigma_A^2 \rho_A & \sigma_A \sigma_B \rho_{AB} \delta \\ \sigma_A \sigma_B \rho_{AB} \delta & \sigma_B^2 \rho_B \end{pmatrix}$$

$$\mathbf{\Lambda} = \begin{pmatrix} \sigma_A^2 (1 - \rho_A) & \sigma_{AB} (1 - \delta) \\ \sigma_{AB} (1 - \delta) & \sigma_B^2 (1 - \rho_B) \end{pmatrix}.$$

$\rho_A$  describe the correlations of measurements made by the method  $A$  at different times. Similarly  $\rho_B$  describe the correlation of measurements made by the method  $B$  at different times. Correlations among repeated measures within the same method are known as intra-class correlation coefficients.  $\rho_{AB}$  describes the correlation of measurements taken at the same same time by both methods. The coefficient  $\delta$  is added for when the measurements are taken at different times, and is a constant of less than 1 for linked replicates. This is based on the assumption that linked replicates measurements taken at the same time would have greater correlation than those taken at different times. For unlinked replicates  $\delta$  is simply 1. Hamlett et al. (2004) provides a useful graphical depiction of the role of each correlation coefficients.

## 1.10 Limits of agreement in LME models

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

$$\text{var}(y_A - y_B) = 2d^2 + \lambda_A^2 + \lambda_B^2. \quad (1.14)$$

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

### 1.10.1 Variance Ratios

The approach proposed by Roy deals with the question of agreement, and indeed interchangeability, as developed by Bland and Altman's corpus of work. In the view of Dunn, a question relevant to many practitioners is which of the two methods is more precise.

The relationship between precision and the within-item and between-item variability must be established. Roy establishes the equivalence of repeatability and within-item variability, and hence precision. The method with the smaller within-item variability can be deemed to be the more precise. A useful approach is to compute the confidence intervals for the ratio of within-item standard deviations (equivalent to the ratio of repeatability coefficients), which can be interpreted in the usual manner. In fact, the ratio of within-item standard deviations, with the attendant confidence interval, can be determined using a single R command: `intervals()`. Pinheiro and Bates (pg 93-95) give a description of how confidence intervals for the variance components are computed. Furthermore a complete set of confidence intervals can be computed to complement the variance component estimates. What is required is the computation of the variance ratios of within-item and between-item standard deviations. A naive approach would be to compute the variance ratios by relevant F distribution quantiles. However, the question arises as to the appropriate degrees of freedom. Limits of agreement are easily computable using the LME framework. While we will not be considering this analysis, a demonstration will be provided in the example.

## 1.11 Testing Procedures

Roy's methodology requires the construction of four candidate models. The first candidate model is compared to each of the three other models successively. It is the alternative model in each of the three tests, with the other three models acting as the respective null models.

The probability distribution of the test statistic can be approximated by a chi-square 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.

Likelihood ratio tests are very simple to implement in R, simply use the 'anova()' commands. Sample output will be given for each variability test. The likelihood ratio test is the procedure used to compare the fit of two models. For each candidate model, the '-2 log likelihood' ( $M2LL$ ) is computed. The test statistic for each of the three hypothesis tests is the difference of the  $M2LL$  for each pair of models. If the  $p$ -value in each of the respective tests exceed as significance level chosen by the analyst, then the null model must be rejected.

$$-2 \ln \Lambda_d = [M2LL \text{ under } H_0 \text{ model}] - [M2LL \text{ under } H_A \text{ model}] \quad (1.15)$$

These test statistics follow a chi-square distribution with the degrees of freedom computed as the difference of the LRT degrees of freedom.

$$\nu = [\text{LRT df under } H_0 \text{ model}] - [\text{LRT df under } H_A \text{ model}] \quad (1.16)$$

## 1.12 Carstensen's Limits of Agreement

Carstensen et al. (2008) presents a methodology to compute the limits of agreement based on LME models. The method of computation is the same as Roy's model, but with the covariance estimates set to zero.

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.

### **1.13 BXC2008 presents - Carstensen’s Limits of agreement**

In cases where there is negligible covariance between methods, the limits of agreement computed using Roy’s model accord with those computed using Carstensen’s model. In cases where some degree of covariance is present between the two methods, the limits of agreement computed using models will differ. In the presented example, it is shown that Roy’s LoAs are lower than those of Carstensen, when covariance is present.

Importantly, estimates required to calculate the limits of agreement are not extractable, and therefore the calculation must be done by hand.

Carstensen et al. (2008) 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 proposing a fitted mixed effects model to obtain appropriate estimates for the variance of the inter-method bias. As their interest lies specifically in extending the Bland-Altman methodology, other formal tests are not considered.

Carstensen et al. (2008) also presents a methodology to compute the limits of agreement based on LME models. The method of computation is similar Roy’s model, but for absence of the covariance estimates. In cases where there is negligible covariance between methods, the limits of agreement computed using Roy’s model accord with

those computed using model described by (??). In cases where some degree of covariance is present between the two methods, the limits of agreement computed using models will differ. In the presented example, it is shown that Roy's LOAs are lower than those of (??), when covariance between methods is present.

Carstensen et al. (2008) 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.

## 1.14 Carstensen's LOAs

Carstensen presents a model where the variation between items for method  $m$  is captured by  $\sigma_m$  and the within item variation by  $\tau_m$ .

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

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

The respective estimates computed by both methods are tabulated as follows. Evidently there is close correspondence between both sets of estimates.

## 1.15 Interaction Terms in Model

Further to Barnhart et al. (2007), if the measurements by a method on an item are not necessarily true replications, e.g., repeated measures over time, then additional terms may be needed for  $e_{mir}$ . ? also addresses this issue by the addition of an interaction term (i.e. a random effect)  $u_{mi}$ , yielding

$$y_{mir} = \alpha_{mi} + u_{mi} + e_{mi}.$$

The additional interaction term is characterized as  $u_{mi} \sim \mathcal{N}(0, \tau_m^2)$  (?).

This extra interaction term provides a source of extra variability, but this variance is not relevant to computing the case-wise differences.

? advises that the formulation of the model should take the exchangeability (in other words, whether or not the measurements are ‘true replicates’) into account. If there is a linkage between measurements (therefore not ‘true’ replicates), the ‘item by replicate’ should be included in the model. If there is no linkage, and the replicates are indeed true replicates, the interaction term should be omitted.

? demonstrates how to compute the limits of agreement for two methods in the case of linked measurements. As a surplus source of variability is excluded from the computation, the limits of agreement are not unduly wide, which would have been the case if the measurements were treated as true replicates.

? also assigns a random effect  $u_{mi}$  for each response  $y_{mir}$ . Importantly Roy’s model assumes linkage.

Carstensen et al. (2008) formulates an LME model, both in the absence and the presence of an interaction term.? uses both to demonstrate the importance of using an interaction term. Failure to take the replication structure into account results in over-estimation of the limits of agreement. For the Carstensen estimates below, an interaction term was included when computed.

Carstensen presents a model where the variation between items for method  $m$  is captured by  $\sigma_m$  and the within item variation by  $\tau_m$ .

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

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

## 1.16 Computation of limits of agreement

The computation thereof require that the variance of the difference of measurements. This variance is easily computable from the variance estimates in the Block -  $\Omega_i$  matrix,

i.e.

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

Carstensen et al. (2008) also presents a methodology to compute the limits of agreement based on LME models. The method of computation is similar Roy's model, but for absence of the covariance estimates. In cases where there is negligible covariance between methods, the limits of agreement computed using Roy's model accord with those computed using model described by (??). In cases where some degree of covariance is present between the two methods, the limits of agreement computed using models will differ. In the presented example, it is shown that Roy's LOAs are lower than those of (??), when covariance between methods is present.

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

The limits of agreement computed by Roy's method are derived from the variance covariance matrix for overall variability. This matrix is the sum of the between subject VC matrix and the within-subject VC matrix. 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.

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

The standard deviation of the differences of methods  $x$  and  $y$  is computed using values from the overall VC matrix.

$$\text{var}(x - y) = \text{var}(x) + \text{var}(y) - 2\text{cov}(x, y)$$

The respective estimates computed by both methods are tabulated as follows. Evidently there is close correspondence between both sets of estimates.

## 1.18 Difference Between Approaches

Carstensen presents two models. One for the case where the replicates, and a second for when they are linked.

Carstensen's model does not take into account either between-item or within-item covariance between methods.

In the presented example, it is shown that Roy's LoAs are lower than those of Carstensen.

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

## 1.19 Differences Between Models

Carstensen et al. (2008) also presents a methodology to compute the limits of agreement based on LME models. In many cases the limits of agreement derived from this method accord with those to Roy's model. However, in other cases dissimilarities emerge. An explanation for this differences can be found by considering how the respective models account for covariance in the observations. Specifying the relevant terms using a bivariate normal distribution, Roy's model allows for both between-method and within-method covariance. Carstensen et al. (2008) formulate a model whereby random effects have univariate normal distribution, and no allowance is made for correlation between observations.

A consequence of this is that the between-method and within-method covariance are zero. In cases where there is negligible covariance between methods, both sets of limits of agreement are very similar to each other. In cases where there is a substantial level of covariance present between the two methods, the limits of agreement computed using models will differ.

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 (??) 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.

## 1.20 Differences

Roy (2009b) 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 method-



ology, the variance of the differences is

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_B^2 + \lambda_A^2 - 2(d_{AB} + \lambda_{AB}) \quad (1.17)$$

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

$$\bar{y}_A - \bar{y}_B \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \quad (1.18)$$

In cases where there is negligible covariance between methods, the limits of agreement computed using Roy's model accord with those computed using Carstensen's model. In cases where some degree of covariance is present between the two methods, the limits of agreement computed using models will differ. In the presented example, it is shown that Roy's LoAs are lower than those of Carstensen, when covariance is present.

Importantly, estimates required to calculate the limits of agreement are not extractable, and therefore the calculation must be done by hand. Carstensen presents a model where the variation between items for method  $m$  is captured by  $\sigma_m$  and the within item variation by  $\tau_m$ .

In contrast to Roy's model, Carstensen's model requires that commonly used 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:  $g_1^2 = g_2^2 = g^2$ . As a consequence, Carstensen's method does not allow for a formal test of the between-item variability.

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

## 1.21 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.

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

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

Roy (2009b) 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.

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  $\hat{\Sigma}$ .

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \tag{1.19}$$

## 1.22 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  $\tau_1 = \tau_2 = \tau$  (Carstensen et al., 2008).*

Another important point is that there is no covariance terms, so further to Carstensen

et al. (2008) 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

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

## 1.23 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.

# Bibliography

- 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. (2010). *Comparing Clinical Measurement Methods : A Practical Guide*. Wiley.
- 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).
- Dunn, G. (2002). *Statistical Evaluation of Measurement Error* (Second ed.). Stanford: American Mathematical Society and Digital Press.
- 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.

- 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. (2009a). An application of linear mixed effects model to assess the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics* 19, 150–173.
- Roy, A. (2009b). 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.