Contents

Chapter 1

LOAS

1.1 Introduction to LME Methods for Computing LoAs

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.

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. 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. In cases where there are repeated measurements by each of the two methods on the same subjects , Bland Altman suggest calculating the mean for each

method on each subject and use these pairs of means to compare the two methods.

The estimate of bias will be unaffected using this approach, but the estimate of the standard deviation of the differences will be too small, because of the reduction of the effect of repeated measurement error. Bland Altman propose a correction for this.

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. Carstensen attends to this issue also, adding that another approach would be to treat each repeated measurement separately. Instead, they recommend a fitted mixed effects model to obtain appropriate estimates for the variance of the inter-method bias. Bendix Carstensen et al. proposed the use of LME models to allow for a more statistically rigourous approach to computing Limits of Agreement. The respective papers also discuss several shortcoming for techniques for dealing with replicate measurements, as proposed by Bland-Altman 1999. Carstensen et al. (2008) computes the limits of agreement to the case with repeated measurements by using LME models. As their interest mainly lies in extending the Bland-Altman methodology, other formal tests are not considered.

The estimate of bias will be unaffected using this approach, but the estimate of the standard deviation of the differences will be too small, because of the reduction of the effect of repeated measurement error. Bland Altman propose a correction for this. Carstensen attends to this issue also, adding that another approach would be to treat each repeated measurement separately.

1.1.1 Carstensen's Mixed Models

Carstensen (2004) and Carstensen et al. (2008) uses an 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. This approach is similar to Deming Regression, similar

to Deming's regression, and for estimating variance components for measurements by different methods.

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 et al. (2008) proposes an approach for comparing two or more methods of measurement based on linear mixed effects models. This approach extends the well established Bland-Altman methodology for the case of replicate measurements on each item. Carstensen considers the matter of computing an appropriate estimate for the standard deviation of case-wise differences, so as to determine the limits of agreement.

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 (1.18). In cases where some degree of covariance is present between the two methods, the limits of agreement computed using models will differ.

Roy (2009b) formulates a very powerful method of assessing the agreement of two methods of measurement, with replicate measurements, also using LME models. Importantly, Roy's approach does not address the issue of limits of agreement (though another related analysis, the coefficient of repeatability, is mentioned).

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. 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.

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.

1.2 Standard Deviation of Differences

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.

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 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 - $\mathbf{\Omega}_i$ matrix, i.e.

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

1.2.1 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 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 - $\boldsymbol{\Omega}_i$ matrix, i.e.

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

1.3 Limits of agreement in LME models

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;

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 λ_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 $\boldsymbol{D}$ and within-subject variability $\boldsymbol{\Lambda}$ can be presented in matrix form,

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

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

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

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

1.3.1 Linked replicates

Carstensen et al. (2008) proposes the addition of an random effects term to their model when the replicates are linked. This term is used to describe the 'item by replicate' interaction, which is independent of the methods. This interaction is a source of variability independent of the methods. Therefore failure to account for it will result in variability being wrongly attributed to the methods.

1.4 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 rth replicate measurement on subject i with method m.

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{1.2}$$

The intercept term α and the $\beta_m \mu_i$ term follow from Dunn (2002), expressing constant and proportional bias respectively , in the presence of a real value μ_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;

$$BLUP_{mir} = \hat{\alpha_m} + \hat{\beta_m}\mu_i + c_{mi} \tag{1.3}$$

. Under the assumption that the μs 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;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad (e_{mi} \sim N(0, \sigma_m^2)) \tag{1.4}$$

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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{1.5}$$

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 τ_1^2 and τ_2^2 separately. Therefore it must be assumed that they are equal. The variance of the difference can be estimated as follows:

$$var(y_{1j} - y_{2j}) \tag{1.6}$$

1.4.1 BXC: extra interaction term

Carstensen (2004) provides an amended formulation which includes an extra interaction term ($d_{mr}d_{mr} \sim N(0, \omega_m^2)$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.

Maximum likelihood estimation is used to estimate the parameters. The REML estimation is not considered since it does not lead to a joint distribution of the estimates of fixed effects and random effects parameters, upon which the assessment of agreement is based.

1.4.2 KP

Most residual covariance structures are design for one within-subject factor. However two or more may be present. For such cases, an approppriate approach would be the residual covariance structure using Kronecker product of the underlying within-subject factor specific covariances structure.

1.5 Carstensen Model (mir model)

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

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2) \tag{1.7}$$

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.

The following model (in the authors own notation) is formulated as follows, where y_{mir} is the rth replicate measurement on subject i with method m.

Using Carstensen's notation, a measurement y_{mi} by method m on individual i the measurement y_{mir} is the rth replicate measurement on the ith item by the mth method, where $m = 1, 2$, $i = 1, \ldots, N$, and $r = 1, \ldots, n_i$ is formulated as follows;

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

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

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \tag{1.9}$$

Here β_0 and β_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 β 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 $\mathrm{E}(b_{mi}) = 0$ with $\mathrm{Var}(b_{mi}) = g_m^2$ and $\mathrm{Cov}(b_{1i}, b_{2i}) = g_{12}$.

The random error term for each response is denoted ϵ_{mir} having $\mathrm{E}(\epsilon_{mir}) = 0$, $\mathrm{Var}(\epsilon_{mir}) = \sigma_m^2$, $\mathrm{Cov}(b_{mir}, b_{m'ir}) = \sigma_{12}$, $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$ and $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{m'ir'}) = 0$.

When two methods of measurement are in agreement, there is no significant differences between β_1 and β_2, g_1^2 andg_2^2, and σ_1^2 andσ_2^2.

Additionally these parameter are assumed to have Gaussian distribution. Two methods of measurement are in complete agreement if the null hypotheses $\mathrm{H}_1\colon \alpha_1 = \alpha_2$ and $\mathrm{H}_2\colon \sigma_1^2 = \sigma_2^2$ and $\mathrm{H}_3\colon g_1^2 = g_2^2$ hold simultaneously. Roy (2009a) uses a Bonferroni correction to control the familywise error rate for tests of $\{\mathrm{H}_1, \mathrm{H}_2, \mathrm{H}_3\}$ and account for difficulties arising due to multiple testing. Additionally, Roy combines H_2 and H_3 into a single testable hypothesis $\mathrm{H}_4\colon \omega_1^2 = \omega_2^2$, where $\omega_m^2 = \sigma_m^2 + g_m^2$ represent the overall variability of method m.

Here the terms α_m and μ_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 ε_{mir}. The c_{mi} term represent random effect parameters corresponding to the two methods, having $\mathrm{E}(c_{mi}) = 0$ with $\mathrm{Var}(c_{mi}) = \tau_m^2$.

Carstensen specifies the variance of the interaction terms as being univariate normally distributed. As such, $\mathrm{Cov}(c_{mi}, c_{m'i}) = 0$. All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method.

The intercept term α and the $\beta_m\mu_i$ term follow from Dunn (2002), expressing constant and proportional bias respectively , in the presence of a real value μ_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).

Of particular importance is terms of the model, a true value for item i (μ_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 presence of the true value term μ_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.

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.

1.6 Using LME models to create Prediction Intervals

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;

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

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.

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

1.7 Prediction Intervals

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 τ_1^2 and τ_2^2 separately. Therefore it must be assumed that they are equal. The variance of the difference can be estimated as follows:

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

1.8 Carstensen's Classical Model

Carstensen's approach is that of a standard two-way mixed effects ANOVA with replicate measurements. With regards to the specification of 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).

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

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

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

The random error term for each response is denoted ε_{mir} having $\mathrm{E}(\varepsilon_{mir}) = 0$, $\mathrm{Var}(\varepsilon_{mir}) = \varphi_m^2$. All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method.

When only two methods are to be compared, separate estimates of τ_m^2 can not be obtained. Instead the average value τ^2 is obtained and used.

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 α can not be estimated, only theiir difference can be estimated as $\bar{D}$

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* (**?**).

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.

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

$$BLUP_{mir} = \hat{\alpha_m} + \hat{\beta_m}\mu_i + c_{mi} \tag{1.13}$$

. Under the assumption that the μs 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) 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. 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.

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.

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 σ_m and the within item variation by τ_m. Further to his model, Carstensen computes the limits of agreement as

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

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 $\boldsymbol{D}$ and $\boldsymbol{\Lambda}$. Using Roy's

methodology, the variance of the differences is

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_A^2 + \lambda_B^2 - 2(d_{AB} + \lambda_{AB}) \tag{1.14}$$

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

$$\bar{y_A} - \bar{y_B} \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \tag{1.15}$$

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,

$$0.0449 \pm 1.96 \times \sqrt{2 \times 0.0596^2 + 0.0772^2 + 0.0724^2} = (-0.220, 0.309). \tag{1.16}$$

1.9 Computing LoAs from LME models

Carstensen presents a model where the variation between items for method m is captured by σ_m and the within item variation by τ_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.

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.

1.9.1 Computation

Modern software packages can be used to fit models accordingly. The best linear unbiased predictor (BLUP) for a specific subject i measured with method m has the

form $BLUP_{mir} = \hat{\alpha_m} + \hat{\beta_m}\mu_i + c_{mi}$, under the assumption that the μs are the true item values.

Maximum likelihood estimation is used to estimate the parameters. The REML estimation is not considered since it does not lead to a joint distribution of the estimates of fixed effects and random effects parameters, upon which the assessment of agreement is based.

This model includes a method by item interaction term.

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.

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

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2) \tag{1.17}$$

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.

1.10 Carstensen 2004 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 can be written as

$$y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + \varepsilon_{mir}. \tag{1.18}$$

The fixed effects α_m and μ_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 $\varepsilon_{mir} \sim \mathcal{N}(0, \varphi_m^2)$. 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. The model expressed in (2) describes measurements by m methods, where $m = \{1, 2, 3 \ldots\}$. Based on the model expressed in (2), Carstensen et al. (2008) compute the limits of agreement as

$$\alpha_1 - \alpha_2 \pm 2\sqrt{\tau_1^2 + \tau_2^2 + \varphi_1^2 + \varphi_2^2}$$

Carstensen et al. (2008) notes that, for $m = 2$, separate estimates of τ_m^2 can not be obtained. To overcome this, the assumption of equality, i.e. $\tau_1^2 = \tau_2^2$ is required.

There is a substantial difference in the number of fixed parameters used by the respective models; the model in (**??**) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items N, whereas the model in (1.18) requires $N + 2$ fixed effects.

Allocating fixed effects to each item i by (1.18) accords with earlier work on comparing methods of measurement, such as Grubbs (1948). However allocation of fixed effects in ANOVA models suggests that the group of items is itself of particular interest, rather than as a representative sample used of the overall population. However this approach seems contrary to the purpose of LOAs as a prediction interval for a population of items. Conversely, Roy (2009a) uses a more intuitive approach, treating the observations as a random sample population, and allocating random effects accordingly.

1.11 Correlation terms

The methodology proposed by Roy (2009b) 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.

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

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.

1.12 Linked replicates

Carstensen et al. (2008) proposes the addition of an random effects term to their model when the replicates are linked. This term is used to describe the 'item by replicate' interaction, which is independent of the methods. This interaction is a source of variability independent of the methods. Therefore failure to account for it will result in variability being wrongly attributed to the methods.

Carstensen et al. (2008) introduces a second data set; the oximetry study. This study done at the Royal Children's Hospital in Melbourne to assess the agreement between co-oximetry and pulse oximetry in small babies.

In most cases, measurements were taken by both method at three different times. In some cases there are either one or two pairs of measurements, hence the data is unbalanced. Carstensen et al. (2008) describes many of the children as being very sick, and with very low oxygen saturations levels. Therefore it must be assumed that a biological change can occur in interim periods, and measurements are not true replicates.

Carstensen et al. (2008) demonstrate the necessity of accounting for linked replicated by comparing the limits of agreement from the 'oximetry' data set using a model with the additional term, and one without. When the interaction is accounted for the limits of agreement are (-9.62,14.56). When the interaction is not accounted for, the limits of agreement are (-11.88,16.83). It is shown that the failure to include this additional term results in an over-estimation of the standard deviations of differences.

Limits of agreement are determined using Roy's methodology, without adding any additional terms, are are found to be consistent with the 'interaction' model; $(-9.562, 14.504)$. Roy's methodology assumes that replicates are linked. However, following Carstensen's example, an addition interaction term is added to the implementation of Roy's model to assess the effect, the limits of agreement estimates do not change. However there is a conspicuous difference in within-subject matrices of both models (denoted 1 and 2 respectively);

```
\begin{equation}
\hat{\boldsymbol{\Lambda}}_{1}= \pmatrix{
16.61 & 11.67\cr
11.67 & 27.65 }\qquad
\boldsymbol{\hat{\Lambda}}_{2}= \pmatrix{
7.55 & 2.60 \cr
2.60 & 18.59}
\end{equation}
```

The variance of the additional random effect in model 2 is 3.01.

The Akaike information criterion (AIC) for both of models are $AIC_1 = 2304.226$ and

$AIC_2 = 2306.226$. Having a difference of AIC values of 2 is equivalent to both models being equally as good as the other. The AIC values for the Carstensen 'unlinked' and 'linked' models are 1994.66 and 1955.48 respectively.

The $\hat{\mathbf{\Lambda}}$ matrices are informative as to the difference between Carstensen's unlinked and linked models. For the oximetry data, the covariance terms (given above as 11.67 and 2.6 respectively) are of similar magnitudes to the variance terms. Conversely for the 'fat' data the covariance term (-0.00032) is negligible. When the interaction term is added to the model, the covariance term remains negligible. (For the 'fat' data, the difference in AIC values is also 2).

To conclude, Carstensen's models provided a rigorous way to determine limits of agreement, but don't provide for the computation of $\hat{\boldsymbol{D}}$ and $\hat{\mathbf{\Lambda}}$. Therefore the test's proposed by Roy (2009a) can not be implemented. Conversely, accurate limits of agreement can be found using Roy's method.

Addition of the interaction term erodes the capability of Roy's methodology to compare candidate models, and therefore shall not be adopted.

(N.B. To complement the blood pressure 'J vs S' analysis, the limits of agreement are $15.62 \pm 1.96 \times 20.33 = (-24.22, 55.46)$.)

1.13 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) $u_m i$, yielding

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

The additional interaction term is characterized as $u_{mi} \sim \mathcal{N}(0, \tau_m^2)$ (Carstensen et al., 2008).

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

Carstensen et al. (2008) 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.

Carstensen et al. (2008) 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 σ_m and the within item variation by τ_m.

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

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

1.14 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.

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

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.

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.

1.15 Carstensen's Limits of agreement

Carstensen et al. (2008) presents a methodology to compute the limits of agreement based on LME models. Importantly, Carstensen's underlying model differs from ARoy2009's

model in some key respects, and therefore a prior discussion of Carstensen's model is required.

1.15.1 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 Carstensen et al. (2008) are very similar to the limits of agreement that follow from the general model.

1.16 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 (1.18) 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 τ_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.

1.16.1 Covariance

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 σ_m and the within item variation by τ_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: $\tau_1^2 = \tau_2^2 = \tau^2$. As a consequence, Carstensen's method does not allow for a formal test of the between-item variability.

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

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.

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

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

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

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

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

1.17 Difference Variance further to Carstensen

Carstensen et al. (2008) states a model where the variation between items for method m is captured by τ_m (our notation d_m^2) and the within-item variation by σ_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 τ_1^2 *and* τ_2^2 *cannot be estimated, only their average value* τ*, so in the case of only two methods we are forced to assume that* $\tau_1 = \tau_2 = \tau$(Carstensen et al., 2008).

1.18 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.

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

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 Ω_i' is the addition of $\hat{D}$ and $\hat{\Sigma}$.

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

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.

Grubbs, F. (1948). On estimating precision of measuring instruments and product variability. *Journal of the American Statistical Association 43*, 243–264.

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.