Contents

0.1 LME Models in Method Comparison Studies

When repeated measures data are available for the the computation of the limits of agreement, it is desirable to use all the data to compare the two methods. The classical Bland-Altman method was developed for two sets of measurements done on one occasion, but is inadequate for replicate measurement data.

Bland and Altman (1999) addressed this issue by suggesting several computationally simple approaches. One approach suggested by Bland and Altman (1999) is to calculate the mean for each method on each subject and use these pairs of means to compare the two methods. Their second approach is to treat each measurement separately. 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 (Carstensen, 2004; Carstensen et al., 2008).

These authors recommends that replicate measurements be used for each method, but recognizes that resulting data are more difficult to analyze. To this end, they recommend the use of LME models as a suitable framework.

Carstensen et al. (2008) extends the well established Bland-Altman methodology for the case of replicate measurements on each itemm by using LME models, to allow for a more statistically rigourous approach to computing appropriate estimates for the variance of the inter-method bias, based upon the variance component estimates derived from the LME models. As their interest mainly lies in extending the Bland-Altman approach, other formal tests are not considered.

0.1.1 Test For Inter-Method Bias in the LME Frameworks

Firstly, a practitioner would investigate whether a significant inter-method bias is present between the methods, as this is the source of disagreement between two methods of measurement that is most easily identified. A formal test for the hypothesis $H_1 : \mu_1 = \mu_2$ can be implemented by examining the fixed effects of the LME model, equivalent to a conventional paired $t-$test. Estimates for the fixed effects yield the

inter-method bias, with the necessary test statistic and $p-$value are typically presented in computer output.

0.1.2   Limits of Agreement in LME models

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 et al remarks that using their approach is common, remarking that "The only slightly non-standard feature is the differing residual variances between methods" (Carstensen, 2010).

Carstensen (2004) proposed LME models for deriving conversion calculations similar to Deming's regression, and for estimating variance components for measurements by different methods. 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;

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

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. The differences are expressed as $d_i = y_{1i} - y_{2i}$.

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

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.

This formulation doesn't require the data set to be balanced, but does require a sufficient number of replicates and measurements to overcome the problem of identifiability. Consequently 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).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.

Carstensen et al. (2008) presents a simplified, but more tractable, model:

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \qquad (3)$$

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.

0.1.3 Carstensen's LME Framework for Method Comparison

Carstensen et al. (2008) recommend a fitted LME model to obtain appropriate estimates for the variance of the inter-method bias. Their interest lies in generalizing the limits-of-agreement (LOA) method developed by Bland and Altman (1986) to take proper cognizance of the replicate measurements. Estimation of repeatability is included in this framework, but other formal tests are not considered

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

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad (e_{mi} \sim N(0, \sigma_m^2)) \qquad (4)$$

For the replicate case, an interaction term c is added to the model, with an associated variance component.

Carstensen (2004) presents a model to describe the relationship between a value of measurement and its real value. The non-replicate case is considered first, as it is the context of the Bland Altman plots. This model assumes that inter-method bias is the only difference between the two methods. A measurement y_{mi} by method m on individual i is formulated as follows:

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad (e_{mi} \sim N(0, \sigma_m^2)). \tag{5}$$

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.

The 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 $m = 1, 2, \ldots, M$ $i = 1, \ldots, N$, and $r = 1, \ldots, n_i$ is formulated as follows;

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + a_{ir} + \epsilon_{mir}, \qquad c_{mi} \sim \mathcal{N}(0, \tau_m^2), a_{ir} \sim \mathcal{N}(0, \varsigma^2), \varepsilon_{mi} \sim \mathcal{N}(0, \varphi_m^2). \tag{6}$$

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + \epsilon_{mir}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \tag{7}$$

Here the terms α_m and μ_i represent the fixed effect for method m and a true value for item i respectively. The β_m term follow from Dunn (2002), expressing constant and proportional bias respectively, in the presence of a real value μ_i. We will just consider the case where $beta = 1$ presently.

0.1.4 Interation Terms

Carstensen et al. (2008) uses both to demonstrate the importance of using an interaction term. Failure to take the replication structure into account results in overestimation of the limits of agreement.

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.

$$y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + \varepsilon_{mir}. \tag{8}$$

The variation between items for method m is captured by σ_m and the within item variation by τ_m.

The random error term for each response is denoted ε_{mir} having $\mathrm{E}(\varepsilon_{mir}) = 0$, $\mathrm{Var}(\varepsilon_{mir}) = \varphi_m^2$. 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 $\varepsilon_{mir} \sim \mathcal{N}(0, \varphi_m^2)$. 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 et al. (2008) specifies the variance of the interaction terms as being univariate normally distributed. As such, $\mathrm{Cov}(c_{mi}, c_{m'i}) = 0$.

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. This model describes measurements by m methods, where $m = \{1, 2, 3 \ldots\}$.

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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \tag{9}$$

The above formulation doesn't require the data set to be balanced. When the design is balanced and there is no ambiguity we can set $n_i = n$. However, the model 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. The quality of exchangeability means that future samples from a population behaves like earlier samples.

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).

0.1.5   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) 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.

0.1.6   Computation of Limits of Agreement in LME models

Carstensen et al. (2008) proposed a technique to calculate prediction intervals in the presence of replicate measurements, overcoming problems associated with Bland-Altman approach in this regard. Between-subject variation for method m is given by d_m^2 (in the author's notation τ_m^2) and within-subject variation is given by σ_m^2.

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. 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, with the average value τ^2 used in practice.

Matrices

The between-subject variability D and within-subject variability Λ can be presented in matrix form,

$$D = \begin{pmatrix} d_A^2 & 0 \\ 0 & d_B^2 \end{pmatrix} = \begin{pmatrix} d^2 & 0 \\ 0 & d^2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \sigma_A^2 & 0 \\ 0 & \sigma_B^2 \end{pmatrix}.$$

The variance for method m is $d_m^2 + \sigma_m^2$. 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

$$\text{var}(y_A - y_B) = 2d^2 + \sigma_A^2 + \sigma_B^2. \tag{10}$$

Importantly the covariance terms in both variability matrices are zero, so no covariance components are present. Further to his model, Carstensen computes the limits of agreement as

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

0.1.7 Carstensen's LOAs

Carstensen et al. (2008) proposes a framework to calculate prediction intervals in the presence of replicate measurements, overcoming problems associated with Bland-Altman approach in this regard. 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. Further to his model, Carstensen computes the limits of agreement as

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

0.1.8 Roy's LME Framework for Method Comparison

Barnhart et al. (2007) sets out three criteria for two methods to be considered in agreement: no significant bias, no difference in the between-subject variabilities, and no significant difference in the within-subject variabilities.

Lack of agreement can also arise if there is a disagreement in overall variabilities. This lack of agreement may be due to differing between-item variabilities, differing within-item variabilities, or both. The formulation previously presented by Roy usefully facilitates a series of significance tests that assess if and where such differences arise. These tests are comprised of a formal test for the equality of between-item variances.

Roy further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal.

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

Roy (2009) proposes a suite of hypothesis tests for assessing the agreement of two methods of measurement, when replicate measurements are obtained for each item, using a LME approach. The tests are implemented by fitting a specific LME model, and three variations thereof, to the data. These three variant models introduce equality constraints that act as null hypothesis cases.

In addition to computing the inter-method bias, three significance tests are carried out on the respective formulations to make a judgement on whether or not two methods are in agreement. The difference in the models are specifically in how the the D and Σ

matrices are constructed, using either an unstructured form or a compound symmetry form. The first model is compared against each of three other models successively. These tests are the pairwise comparison of candidate models, one formulated without constraints, the other with a constraint.

Four candidates models are fitted to the data. These models are similar to one another, but for the imposition of equality constraints. The tests are implemented by fitting a four variants of a specific LME model to the data. For the purpose of comparing models, one of the models acts as a reference model while the three other variant are nested models that introduce equality constraints to serves as null hypothesis cases. The framework uses a linear mixed effects regression fit using a combination of symmetric and compound symmetry (CS) correlation structure the variance covariance matrices.

Roy (2009) uses the same definition of replicate measurement as Bland and Altman (1999); measurements taken in quick succession by the same observer using the same instrument on the same subject can be considered true replicates under identical conditions. Roy (2009) notes that some measurements may not be 'true' replicates, as data can not be collected in this way. In such cases, the correlation matrix on the replicates may require a different structure, such as the autoregressive order one $AR(1)$ structure. However determining MLEs with such a structure would be computational intense, if possible at all.

For these tests, four candidate LME models are constructed. The differences in the models are specifically in how the the D and Λ matrices are constructed, using either an unstructured form or a compound symmetry form. To illustrate these differences, consider a generic matrix A,

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

A symmetric matrix allows the diagonal terms a_{11} and a_{22} to differ. The compound symmetry structure requires that both of these terms be equal, i.e $a_{11} = a_{22}$.

0.1.9 Model Specification for Roy's Hypotheses Tests

The LME model underpinning Roy's approach can be written

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \tag{11}$$

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 $E(b_{mi}) = 0$ with $Var(b_{mi}) = g_m^2$ and $Cov(b_{1i}, b_{2i}) = g_{12}$. The random error term for each response is denoted ϵ_{mir} having $E(\epsilon_{mir}) = 0$, $Var(\epsilon_{mir}) = \sigma_m^2$, $Cov(\epsilon_{1ir}, \epsilon_{2ir}) = \sigma_{12}$, $Cov(\epsilon_{mir}, \epsilon_{mir'}) = 0$ and $Cov(\epsilon_{1ir}, \epsilon_{2ir'}) = 0$. Additionally these parameter are assumed to have Gaussian distribution. Two methods of measurement are in complete agreement if the null hypotheses $H_1: \alpha_1 = \alpha_2$ and $H_2: \sigma_1^2 = \sigma_2^2$ and $H_3: g_1^2 = g_2^2$ hold simultaneously. Roy (2009) uses a Bonferroni correction to control the familywise error rate for tests of $\{H_1, H_2, H_3\}$ and account for difficulties arising due to multiple testing. Additionally, Roy combines H_2 and H_3 into a single testable hypothesis $H_4: \omega_1^2 = \omega_2^2$, where $\omega_m^2 = \sigma_m^2 + g_m^2$ represent the overall variability of method m.

Response for ith subject can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + b_{1i} z_{i1} + b_{2i} z_{i2} + \epsilon_i$$

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

0.1.10 Formulation of the Response Vector

Information of individual i is recorded in a response vector y_i. The response vector is constructed by stacking the response of the 2 responses at the first instance, then the 2 responses at the second instance, and so on. Therefore the response vector is a $2n_i \times 1$ column vector. The covariance matrix of y_i is a $2n_i \times 2n_i$ positive definite matrix Ω_i.

Consider the case where three measurements are taken by both methods A and B, y_i is a 6×1 random vector describing the ith subject.

$$y_i = (y_i^{A1}, y_i^{B1}, y_i^{A2}, y_i^{B2}, y_i^{A3}, y_i^{B3})\prime$$

The response vector y_i can be formulated as an LME model according to Laird-Ware form.

$$y_i = X_i\beta + Z_ib_i + \epsilon_i$$
$$b_i \sim \mathcal{N}(0, D)$$
$$\epsilon_i \sim \mathcal{N}(0, R_i)$$

Information on the fixed effects are contained in a three dimensional vector $\beta = (\beta_0, \beta_1, \beta_2)\prime$. For computational purposes β_2 is conventionally set to zero. Consequently β is the solutions of the means of the two methods, i.e. $E(y_i) = X_i\beta$. The variance covariance matrix D is a general 2×2 matrix, while R_i is a $2n_i \times 2n_i$ matrix.

0.1.11 Likelihood Ratio Tests

The first model acts as an alternative hypothesis to be compared against each of three other models, acting as null hypothesis models, successively. The models are compared using the likelihood ratio test.

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 models are compared using the likelihood ratio test, a general method for comparing nested models fitted by ML (**?**).

Likelihood ratio tests are a class of tests based on the comparison of the values of the likelihood functions of two candidate models. LRTs can be used to test hypotheses about covariance parameters or fixed effects parameters in the context of LMEs. The test statistic for the likelihood ratio test is the difference of the log-likelihood functions, multiplied by -2. The probability distribution of the test statistic is approximated by

the χ^2 distribution with $(\nu_1 - \nu_2)$ degrees of freedom, i.e. the difference between the degrees of freedom of models 1 and 2 respectively.

0.1.12   Roy's Tests of Variances

Roy (2009) proposes a series of three tests on the variance components of an LME model. For these tests, four candidate models are fitted to the data, each differing by various constraints applied to the variance covariance matrices.

Three tests of hypothesis are provided, appropriate for evaluating the agreement between the two methods of measurement under this sampling scheme.

Variability Test 1

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

$$H_0 : \ d_1 = d_2$$
$$H_A : \ d_1 \neq d_2$$

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

The first test allows of the comparison the begin-subject variability of two methods. Similarly, the second test assesses the within-subject variability of two methods. A third test is a test that compares the overall variability of the two methods. Other important aspects of the method comparison study are consequent. The limits of agreement are computed using the results of the first model.

$$H_3 : \sigma_1^2 = \sigma_2^2$$
$$K_3 : \sigma_1^2 \neq \sigma_2^2$$

Variability Test 2

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

$$H_0 : \ \sigma_1 = \sigma_2$$
$$H_A : \ \sigma_1 \neq \sigma_2$$

This model is performed in the same manner as the first test, only reversing the roles of $l\hat{D}$ and $l\hat{\Sigma}$. The null model is constructed a symmetric form for $\hat{\Sigma}$ while the alternative model uses a compound symmetry form. This time $l\hat{D}$ has a symmetric form for both models, and will be the same for both.

As the within-subject variabilities are fundamental to the coefficient of repeatability, this variability test likelihood ratio test is equivalent to testing the equality of two coefficients of repeatability of two methods. In presenting the results of this test, Roy (2009) includes the coefficients of repeatability for both methods.

Variability Test 3

Roy also integrates H_2 and H_3 into a single testable hypothesis $H_4 \colon \omega_1^2 = \omega_2^2$, where $\omega_m^2 = \sigma_m^2 + d_m^2$ represent the overall variability of method m. Roy (2009) further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal. An examination of this topic is useful because a method for computing Limits of Agreement follows from here.

A formal test for the equality of overall variances is also presented.

$$H_4 : \omega_1^2 = \omega_2^2$$
$$K_4 : \omega_1^2 \neq \omega_2^2$$

Two methods can be considered to be in agreement if criteria based upon these techniques are met. Additionally Roy makes reference to the overall correlation coefficient of the two methods, which is determinable from variance estimates.

Conversely, the tests of variability required detailed explanation. Each test is performed by fitting two candidate models, according with the null and alternative hypothesis respectively. The distinction between the models arise in the specification in one, or both, of the variance-covariance matrices.

Disagreement in overall variability may be caused by different between-item variabilities, by different within-item variabilities, or by both. If the exact cause of disagreement between the two methods is not of interest, then the overall variability test H_4 is an alternative to testing H_2 and H_3 separately.

The estimated overall variance covariance matrix 'Block Ω_i' is the addition of estimate of the between-subject variance covariance matrix $\hat{D}$ and the within-subject variance covariance matrix $\hat{\Sigma}$.

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \tag{12}$$

Overall variability between the two methods (Ω) is sum of between-subject (D) and within-subject variability (Σ), Roy (2009) denotes the overall variabilityas Block - Ω_i. The overall variation for methods 1 and 2 are given by

$$\text{Block } \Omega_i = \begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix} = \begin{pmatrix} d_1^2 & d_{12} \\ d_{12} & d_2^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

The last of the variability test examines whether or not both methods have the same overall variability. This enables the joint consideration of second and third criteria.

$$H_0: \ \omega_1 = \omega_2$$
$$H_A: \ \omega_1 \neq \omega_2$$

The null model is constructed a symmetric form for both $\hat{D}$ and $\hat{\Lambda}$ while the alternative model uses a compound symmetry form for both.

0.1.13 Computing Limits of Agreement Using Roy's Model

Roy (2009) has demonstrated a method whereby d_A^2 and d_B^2 can be estimated separately. Also covariance terms are present in both D and Σ. Using Roy's approach, the variance of case-wise difference in measurements can be determined from Block-Ω_i. Hence limits of agreement can be computed. The computation of the limits of agreement require that the variance of the difference of measurements. This variance is easily computable from the estimate of the Block - Ω_i matrix. The variance of differences is easily computable from the variance estimates in the Block - Ω_i matrix, i.e.

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

Lack of agreement can arise if there is a disagreement in overall variabilities.

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 Variance Covariance matrix. 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); $0.044884 \pm 1.96 \times 0.1373979 = (-0.224, 0.314)$.

0.2 Differences Between Approaches

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.

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. Therefore the variance covariance matrices for between-item and within-item variability are respectively.

$$
D = \begin{pmatrix} d_2^1 & 0 \\ 0 & d_2^2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma_2^1 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}
$$

As a consequence, Carstensen's method does not allow for a formal test of the between-item variability.

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 Carstensen et al. (2008).

A consequence of this is that the between-method and within-method covariance are 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.

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

There is a substantial difference in the number of fixed parameters used by the respective models; the model in Roy (2009) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items N, whereas the model using the Carstensen Model requires $N + 2$ fixed effects. Allocating fixed effects to each item i using Carstensen's model 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 (2009) uses a more intuitive approach, treating the observations as a random sample population, and allocating random effects accordingly.

0.2.1 Comparing MCS Approaches

Roy's tests afford the opportunity to expand upon Carstensen's approach.

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.

Roy (2009) formulated a very powerful method of assessing the agreement of two methods of measurement, with replicate measurements, also using LME models. This approach does not directly address the issue of limits of agreement, but does allow for an alternative approach to computing LoAs using LME Models.

There is a substantial difference in the number of fixed parameters used by the respective models; the model in (11) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items N, whereas the model in (8) requires $N + 2$ fixed effects.

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.

The presence of the true value term μ_i gives rise to an important difference between Carstensen's and Roys'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.

Allocating fixed effects to each item i by (8) 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 (2009) uses a more intuitive approach, treating the observations as a random sample population, and allocating random effects accordingly.

Finally, to complement the blood pressure (i.e.'J vs S') method comparison from the previous section (i.e.'J vs S'), the limits of agreement are $15.62 \pm 1.96 \times 20.33 = (-24.22, 55.46)$.)

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 Roys's LoAs are lower than those of Carstensen.

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

Carstensen's approach is that of a standard two-way mixed effects ANOVA with replicate measurements.

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.

0.2.2 Summary

Carstensen et al. (2008) remark that modern statistical computation, such as that used for LME models, greatly improve the efficiency of calculation compared to previous 'by-hand' approaches, as advocated in Bland and Altman (1999), describing them as tedious, unnecessary and outdated. Instead estimates for required LME parameters can be read directly from program output. Furthermore, using computer approaches removes associated constraints, such as the need for the design to be perfectly balanced.

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 methods of measurement: Extending the loa by regression. *Statistics in medicine 29*(3), 401–410.

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.

Lam, M., K. Webb, and D. O'Donnell (1999). Correlation between two variables in repeated measurements. *American Statistical Association, Proceedings of the Biometric Session*, 213–218.

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