Contents

Chapter 1

LME Models in Method Comparison Studies

1.1 LME Models in Method Comparison Studies

When repeated measurements are available for calculating the limits of agreement, it is desirable to use all available 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 proposed approach 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).

Carstensen et al. (2008) 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, a view shared by

Roy (2009) etc.

Carstensen et al. (2008) extends the well established Bland-Altman methodology for the case of replicate measurements on each item 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.

f

1.1.1 Computing Limits of Agreement with LME Methods

Carstensen (2004) and Carstensen et al. (2008) uses an LME model to compute limits of agreement where replicate measurements are available on each item, proposing an approach for comparing two or more methods of measurement based on linear mixed effects models.

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, by computing an appropriate estimate for the standard deviation of case-wise differences, so as to determine the limits of agreement. Carstensen et al. (2008) took issue with the limits of agreement based on mean values of replicate measurements, in that they can only be interpreted as prediction limits for difference between means of repeated measurements by both methods, as opposed to the difference of all measurements. Incorrect conclusions would be caused by such a misinterpretation.

1.1.2 Carstensen's LME Framework for Method Comparison

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

Carstensen et al. (2008) recommend a fitted LME model to obtain appropriate estimates for the variance of the inter-method bias. Estimation of repeatability is included in this framework, but other formal tests are not considered.

Carstensen et al. (2008) develop this approach is that of a standard two-way mixed effects ANOVA with replicate measurements, reformulated for the case of replicate measurements, with random effects terms specified as appropriate. With regards to specifying the variance terms, Carstensen et al. (2008) remarks that using their 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 (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:

$$y_{mi} = \alpha_m + \beta_m \mu_i + \epsilon_{mi}. \tag{1.1}$$

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. The random error term for each response is denoted ϵ_{mi} with $\text{E}(\epsilon_{mi}) = 0$ and $\text{Var}(\epsilon_{mi}) = \sigma^2_m$.

1.1.3 Test For Inter-Method Bias in the LME Frameworks

Two methods can be considered to be in agreement if criteria based upon these tests are met. 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 $\mathrm{H}_1 : \mu_1 = \mu_2$ 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, typically accompanied by the necessary test statistic and $p-$value in programming software output. Inference for inter-method bias follows from well-established methods and, as such, will only be noted when describing examples.

Replicate Measurement Case

Carstensen et al. (2008) proposes the addition of an random effects term to their model when the replicates are linked. For the replicate case, an interaction term c is added to the model, with an associated variance component. This term is used to describe the 'item by replicate' interaction, a_{ir}, 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) 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.

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;

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + a_{ir} + \epsilon_{mir}, \tag{1.2}$$

The random error term for each response is denoted ϵ_{mir} with $\mathrm{E}(\epsilon_{mir}) = 0$ and $\mathrm{Var}(\epsilon_{mir}) = \sigma^2_m$. Between-subject variation for method m is given by d^2_m (in the author's notation τ^2_m) and within-subject variation is given by σ^2_m.

The random-effect terms comprise an item-by-replicate interaction term $a_{ir} \sim \mathcal{N}(0, \sigma^2)$, a method-by-item interaction term $c_{mi} \sim N(0, \tau^2_m)$. ϵ_{mir} is the residual associated with each observation, with $\varepsilon_{mir} \sim N(0, \varphi^2_m)$.

c_{mi} is a interaction term to account for replicate, and ϵ_{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 c_{mi} term represent random effect parameters corresponding to the two methods, having $\mathrm{E}(c_{mi}) = 0$ with $\mathrm{Var}(c_{mi}) = \tau^2_m$. Carstensen et al. (2008) specifies the variance of the interaction terms as being univariate normally distributed. As such, $\mathrm{Cov}(c_{mi}, c_{m'i}) = 0$.

This 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.

Simplified Model

All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method. For the case when replicate measurements are assumed to be exchangeable for item i, a_{ir} can be removed.Equation 1.2 can be re-expressed as

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + \epsilon_{mir}, \tag{1.3}$$

with all of the component terms defined as they were before.

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.

1.1.4 Computation of Limits of Agreement in LME models

Carstensen et al. (2008) demonstrates how to compute the limits of agreement for two methods in the case of linked measurements. By specifying a model with interaction terms, surplus sources of variability are excluded from the computation. Consequently the limits of agreement are not unduly wide, which would have been the case if the measurements were treated as true replicates.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + \epsilon_{mir}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma^2_m), \quad c_{mi} \sim \mathcal{N}(0, \tau^2_m). \tag{1.4}$$

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

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

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.

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. 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 $d_i = y_{1i} - y_{2i}$. Carstensen et al. (2008) compute the limits of agreement as

$$\hat{\alpha}_1 - \hat{\alpha}_2 \pm \pm 2\sqrt{\tau^2_1 + \tau^2_2 + \varphi^2_1 + \varphi^2_2} \tag{1.6}$$

Carstensen et al. (2008) states a model where the variation between items for method m is captured by τ_m (our notation d^2_m) and the within-item variation by σ_m. Carstensen et al. (2008) notes that, for $m = 2$, separate estimates of τ^2_m can not be obtained. When only two methods are compared, Carstensen et al. (2008) notes that separate estimates of τ^2_m 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. Equation1.6 can be re-epxressed as follows

$$\hat{\alpha}_1 - \hat{\alpha}_2 \pm \sqrt{2\hat{\tau}^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2} \tag{1.7}$$

Importantly the covariance terms in both variability matrices are zero, so no covariance components are present.

Matrices

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

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

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

$$\text{var}(y_A - y_B) = 2d^2 + \sigma_A^2 + \sigma_B^2. \tag{1.8}$$

1.2 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 may be identified by disagreement in overall variabilities. This lack of agreement may be due to differing between-item variabilities, differing within-item variabilities, or both. 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 formulation previously presented by Roy (2009) 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.

For the purposes of comparing two methods of measurement, Roy (2009) presents a framework utilizing LME models. 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 covariance matrices are constructed, using either an unstructured form or a compound symmetry form. This is a key component of this testing process.

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) 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.

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.

Four candidates models are fitted to the data. 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.

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.

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

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 $a_{11} = a_{22}$.

1.2.1 Model Specification for Roy's Hypotheses Tests

The LME model underpinning Roy's approach can be written

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

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 $\text{E}(b_{mi}) = 0$ with $\text{Var}(b_{mi}) = g^2_m$ and $\text{Cov}(b_{1i}, b_{2i}) = g_{12}$. The random error term for each response is denoted ϵ_{mir} having $\text{E}(\epsilon_{mir}) = 0$, $\text{Var}(\epsilon_{mir}) = \sigma^2_m$, $\text{Cov}(\epsilon_{1ir}, \epsilon_{2ir}) = \sigma_{12}$, $\text{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$ and $\text{Cov}(\epsilon_{1ir}, \epsilon_{2ir'}) = 0$. Additionally these parameter are assumed to have Gaussian distribution. Two methods of measurement are in complete agreement if the null hypotheses

$\mathrm{H}_1\colon \alpha_1 = \alpha_2$ 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 (2009) 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.

1.2.2 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 $\boldsymbol{\Omega}_i$.

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 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 $y_i = (y_{1i1}, y_{2i1}, y_{1i2}, \ldots, y_{mir}, \ldots, y_{1in_i}, y_{2in_i})'$. The response vector y_i can be formulated as an LME model according to classical Laird-Ware form.

$$y_i = X_i\beta + Z_i b_i + \epsilon_i, \tag{1.10}$$

with $b_i \sim \mathcal{N}(0, \boldsymbol{G})$ and $\epsilon_i \sim \mathcal{N}(0, \boldsymbol{R_i})$. 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 $\boldsymbol{G}$ is a general 2×2 matrix, while $\boldsymbol{R}_i$ is a $2n_i \times 2n_i$ matrix.

1.2.3 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, a general method for comparing nested models fitted by ML (**?**) *Lehmann (2006).*

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 (**?**) *Lehmann (2006).*

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.

1.2.4 Roy's Tests of Variances

Roy (2009) proposes a series of three tests on the variance components of an LME model, appropriate for evaluating the agreement between the two methods of measurement under this sampling scheme. 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.

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.

Joint consideration of second and third criteria is enabled by a formal test for the equality of overall variances, $\mathrm{H}_4 : \omega_1^2 = \omega_2^2$ is also presented. 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.

Variability Test 1

The first test determines of $H_2 : d_1 = d_2$, whether or not both methods A and B have the same between-subject variability, further to the second of Roy's criteria.

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 $\hat{\boldsymbol{G}}$ (i.e. the null model). For this test $\hat{\boldsymbol{\Sigma}}$ 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.

Variability Test 2

This test determines whether or not both methods have the same within-subject variability, i.e. $H_0 : \sigma_1 = \sigma_2$ thus enabling a decision on the third of Roy's criteria.

This model is performed in the same manner as the first test, only reversing the roles of $\hat{\boldsymbol{G}}$ and $\hat{\boldsymbol{\Sigma}}$. The null model is constructed a symmetric form for $\hat{\Sigma}$ while the alternative model uses a compound symmetry form. This time $\hat{\boldsymbol{G}}$ 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 (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.

Roy (2009) denotes the overall variability as Block-Ω_i, defining it as the addition of estimate of the between-subject variance covariance matrix $\hat{G}$ and the within-subject variance covariance matrix $\hat{\Sigma}$, i.e. Block $\Omega_i = \hat{G} + \hat{\Sigma}$,

$$\text{Block } \Omega_i = \begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix} = \begin{pmatrix} d_1^2 & d_{12} \\ d_{12} & d_2^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \tag{1.11}$$

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

1.2.5 Computing Limits of Agreement Using Roy's Model

Roy (2009) has demonstrated a method whereby d^2_A and d^2_B can be estimated separately. Also covariance terms are present in both $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$. Using Roy's approach, the variance of case-wise difference in measurements can be determined from Block-$\boldsymbol{\Omega}_i$. Hence limits of agreement can be computed. The computation of the limits of agreement require that the variance of the difference of measurements. This variance is easily computable from the estimate of the Block - Ω_i matrix. The variance of differences is easily computable from the variance estimates in the Block - Ω_i matrix, i.e.

$$\text{Var}(y_1 - y_2) = \sqrt{\omega^2_1 + \omega^2_2 - 2\omega_{12}}. \tag{1.12}$$

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); $0.045 \pm 1.96 \times 0.137 = (-0.224, 0.314)$.

1.2.6 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's tests are complemented by the ability to the overall correlation coefficient of the two methods, which are estimable from variance estimates.

In addition to the variability tests, Roy (2009) 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, and demonstrates that placing undue importance to it can lead to incorrect conclusions.

Two methods can be considered to be in agreement if criteria based upon these tests are met. Inference for inter-method bias follows from well-established methods and, as such, will only be noted when describing examples.

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

$$\text{Block-}\boldsymbol{\Omega_i} = \hat{\boldsymbol{G}} + \hat{\boldsymbol{\Sigma}} \tag{1.13}$$

Overall variability between the two methods (Ω) is sum of between-subject (D) and within-subject variability ($\boldsymbol{\Sigma}$), Roy (2009) denotes the overall variability as Block-Ω_i. The overall variation for methods 1 and 2 are given by

$$\text{Block } \Omega_i = \left(\begin{array}{cc} \omega^2_1 & \omega_{12} \\ \omega_{12} & \omega^2_2 \end{array}\right) = \left(\begin{array}{cc} d^2_1 & d_{12} \\ d_{12} & d^2_2 \end{array}\right) + \left(\begin{array}{cc} \sigma^2_1 & \sigma_{12} \\ \sigma_{12} & \sigma^2_2 \end{array}\right) \tag{1.14}$$

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

1.3 Correlation

Roy (2009) remarks that current computer implementations (e.g. PROC MIXED) only gives overall correlation coefficients, but not their variances. Similarly variance are not determinable in `R` as yet either. Consequently it is not possible to carry out inferences based on all overall correlation coefficients.

1.3.1 Formal Testing for Covariances

The within-item variability is specified as follows, where x and y are the methods of measurement in question.

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

σ_x^2 and σ_y^2 describe the level of measurement error associated with each of the measurement methods for a given item. Attention must be given to the off-diagonal elements of the matrix.

It is intuitive to consider the measurement error of the two methods as independent of each other. A formal test can be performed to test the hypothesis that the off-diagonal terms are zero.

As it is pertinent to the difference between the two described methodologies, the facilitation of a formal test would be useful. Extending the approach proposed by Roy (2009), the test for overall covariance can be formulated as $\mathrm{H}_5 : \sigma_{xy} = 0$.

As with the tests for variability, this test is performed by comparing a pair of model fits corresponding to the null and alternative hypothesis. In addition to testing the overall covariance, similar tests can be formulated for both the component variabilities if necessary.

1.4 Extension of Roy's Technique

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.

1.4.1 Application of Roy's Approach For Non-Replicate Measurements

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.

1.5 Differences Between Approaches

Both Carstensen et al. (2008) and Roy (2009) present methodologies to compute the limits of agreement based on LME models. In many cases the limits of agreement derived from this method accord with each other. However, in other cases dissimilarities emerge.

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.

The presence of the true value term μ_i gives rise to an important difference between these approaches. 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; the model in (1.9) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items N.

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.

Conversely the model in (1.5) requires $N+2$ fixed effects. Allocating fixed effects to each item i by (1.5) 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.

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.

Arguably, Roy (2009) uses a more intuitive approach, treating the observations as a random sample population, and allocating random effects accordingly. In other words, this model 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.

Another important difference is that (**??**) 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. Therefore the variance covariance matrices for between-item and within-item variability are respectively.

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.

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.

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).

Conversely, Roy (2009) uses a more intuitive approach, treating the observations as a random sample population, and allocating random effects accordingly. 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.

$$G = \begin{pmatrix} d^1_2 & 0 \\ 0 & d^2_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma^1_2 & 0 \\ 0 & \sigma^2_2 \end{pmatrix}$$

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

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 model described by Carstensen et al. (2008). 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.

$$\begin{pmatrix} \omega_1^2 & 0 \\ 0 & \omega_2^2 \end{pmatrix} = \begin{pmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

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 cases where there is negligible covariance between methods, the limits of agreement computed using 1.9 accord with those computed using 1.5. In cases where some degree of covariance is present between the two methods, the limits of agreement computed will differ between models. In the presented example, it is shown that Roy's LOAs are lower than those of Carstensen, when covariance is present.

Finally, implementation requires that the between-item variances are estimated as the same value: $d_1^2 = d_2^2 = d^2$. Necessarily Carstensen's method does not allow for a formal test of the between-item variability.

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

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

$$\begin{pmatrix} \omega_1^2 & 0 \\ 0 & \omega_2^2 \end{pmatrix} = \begin{pmatrix} d^2 & 0 \\ 0 & d^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

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.

1.5.1 Comparing MCS 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 Roys's LoAs are lower than those of Carstensen.

$$\begin{pmatrix} \omega^1_2 & 0 \\ 0 & \omega^2_2 \end{pmatrix} = \begin{pmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{pmatrix} + \begin{pmatrix} \sigma^2_1 & 0 \\ 0 & \sigma^2_2 \end{pmatrix}$$

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

1.6 Conclusion

Carstensen et al. (2008) and Roy (2009) highlight the need for method comparison methodologies suitable for use in the presence of replicate measurements. Roy (2009) presents a comprehensive methodology for assessing the agreement of two methods, for replicate measurements. This methodology has the added benefit of overcoming the problems of unbalanced data and unequal numbers of replicates.

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

Implementation of the methodology, and interpretation of the results, is relatively easy for practitioners who have only basic statistical training. Furthermore, it can be shown that widely used existing methodologies, such as the limits of agreement, can be incorporated into Roy's methodology.

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.

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.

Furthermore, it can be shown that widely used existing methodologies, such as the limits of agreement, can be incorporated into Roy's methodology.

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 $15.62 \pm 1.96 \times 20.33 =$

$(-24.22, 55.46)$.)

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

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.

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. (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.