Contents

Chapter 1

Linear Mixed Effects Models

In cases where there are repeated measurements by each of the two methods on the same subjects, Bland and Altman (1999) 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 incorrect (Carstensen, 2004). Carstensen (2004) recommends that replicate measurements for each method, but recognizes that resulting data are more difficult to analyze. To this end, Carstensen (2004) and Carstensen et al. (2008) recommend the use of LME models as a suitable framework for method comparison in the case of repeated measurements. Roy (2009) uses an LME model approach to provide a set of formal tests for method comparison studies.

Several authors, such as Kelly (1985) and **?**, recommend the use of structural equation models for method comparison. This approach provides a statistically rigourous analysis, but the approach is undermined in several ways. LME models have greater flexibility and can be adapted to any variant of the method comparison research question, whereas SEM is suitable for specific questions only. Highly complex models can be developed using SEM, but to overcome the problem of identifiability, a large quantity of data must be gathered. Often this is beyond what is practical in the main applications of method comparison studies, namely the medical sciences. Once simplifications are applied, there is little functional difference between SEM and LMEs.

1.1 Linear Mixed Effects Models

A linear mixed effects (LME) model is a statistical model containing both fixed effects and random effects (also known as variance components). LME models are a generalization of the classical linear model, which contain fixed effects only. When the levels of factors are considered to be sampled from a population, and each level is not of particular interest, they are considered random quantities with associated variances. The effects of the levels, as described, are known as random effects. Random effects are represented by unobservable normally distributed random variables. Conversely fixed effects are considered non-random and the levels of each factor are of specific interest.

Fisher (1918) introduced variance components models for use in genetical studies. Whereas an estimate for variance must take an non-negative value, an individual variance component, i.e. a component of the overall variance, may be negative.

The LME framework has developed since, including contributions from Tippett (1931), who extend the use of variance components into linear models, and Eisenhart (1947), who introduced the 'mixed model' terminology and formally distinguished between mixed and random effects models. **?** devised a methodology for deriving estimates for both the fixed effects and the random effects, using a set of equations that would become known as 'mixed model equations' or 'Henderson's equations'.

LME modelling was further enhanced by Henderson's later works (Henderson, 1953; Henderson et al., 1959, 1963, 1973, 1984). The key features of Henderson's work provide the basis for the estimation techniques.

Hartley and Rao (1967) demonstrated that unique estimates of the variance components could be obtained using maximum likelihood methods. However these estimates are known to be biased 'downwards' (i.e. underestimated), because of the assumption that the fixed estimates are known, rather than being estimated from the data. Patterson and Thompson (1971) produced an alternative set of estimates, known as the restricted maximum likelihood (REML) estimates, that do not require the fixed effects to be known. Thusly there is a distinction the REML estimates and the original estimates, now commonly referred to as ML estimates.

1.1.1 Laird-Ware Model

Laird and Ware (1982) provides a form of notation for notation for LME models that has since become the standard form, or the basis for more complex formulations. Due to computation complexity, linear mixed effects models have not seen widespread use until many well known statistical software applications began facilitating them. SAS Institute added PROC MIXED to its software suite in 1992 (Singer, 1998). Pinheiro and Bates (1994) described how to compute LME models in the `S-plus` environment.

Linear mixed effects models (LME) differs from the conventional linear model in that it has both fixed effects and random effects regressors, and coefficients thereof. The notation provided here is generic, and will be adapted to accord with complex formulations that will be encountered in due course. Using Laird-Ware form, the LME model is commonly described in matrix form,

$$Y = X\beta + Zb + \epsilon \tag{1.1}$$

Y is the $n \times 1$ response vector, where n is the number of observations. β is a $p \times 1$ vector of fixed p effects, with the first element being the population mean. X and Z are $n \times p$ and $n \times q$ “model matrices“ for fixed effects and random effects respectively, comprising 0s or 1s, depending on the observation is question. The vector of residuals, $v(e)$ has dimension $n \times 1$. The random effects are contained in the $q \times 1$ vector $\mathbf{b}$.

1.1.2 LME Model Estimation

Estimation of LME models involve two complementary estimation issues’; estimating the vectors of the fixed and random effects estimates $\hat{\beta}$ and $\hat{b}$ and estimating the variance covariance matrices G and Σ. Inference about fixed effects have become known as ‘estimates’, while inferences about random effects have become known as ‘predictions’. The most common approach to obtain estimators are Best Linear Unbiased Estimator

(BLUE) and Best Linear Unbiased Predictor (BLUP). For an LME model given by (1.1), the BLUE of $\hat{\beta}$ is given by

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y,$$

whereas the BLUP of $\hat{b}$ is given by

$$\hat{b} = DZ'V^{-1}(y - X\hat{\beta}).$$

Henderson's Equations

Because of the dimensionality of V (i.e. $n \times n$) computing the inverse of V can be difficult. As a way around the this problem Henderson (1953); Henderson et al. (1959, 1963, 1973, 1984) offered a more simpler approach of jointly estimating $\hat{\beta}$ and $\hat{b}$. **?** made the (ad-hoc) distributional assumptions $y|b \sim \mathrm{N}(X\beta + Zb, \Sigma)$ and $b \sim \mathrm{N}(0, G)$, and proceeded to maximize the joint density of y and b

$$\begin{vmatrix} G & 0 \\ 0 & \Sigma \end{vmatrix}^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix}'\begin{pmatrix} G & 0 \\ 0 & \Sigma \end{pmatrix}^{-1}\begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix}\right\}, \quad (1.2)$$

with respect to β and b, which ultimately requires minimizing the criterion

$$(y - X\beta - Zb)'\Sigma^{-1}(y - X\beta - Zb) + b'G^{-1}b. \quad (1.3)$$

This leads to the mixed model equations

$$\begin{pmatrix} X'\Sigma^{-1}X & X'\Sigma^{-1}Z \\ Z'\Sigma^{-1}X & X'\Sigma^{-1}X + D^{-1} \end{pmatrix}\begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X'\Sigma^{-1}y \\ Z'\Sigma^{-1}y \end{pmatrix}. \quad (1.4)$$

Using these equations, obtaining the estimates requires the inversion of a matrix of dimension $p + q \times p + q$, considerably smaller in size than V. Henderson et al. (1963) shows that these mixed model equations do not depend on normality and that $\hat{\beta}$ and $\hat{b}$ are the BLUE and BLUP under general conditions, provided G and Σ are known.

Robinson (1991) points out that although **?** initially referred to the estimates $\hat{\beta}$ and $\hat{b}$ from (1.4) as "joint maximum likelihood estimates", Henderson (1973) later advised that these estimates should not be referred to as "maximum likelihood" as the function being maximized in (1.3) is a joint density rather than a likelihood function. Lee et al. (2006) remarks that it is clear that Henderson used joint estimation for computational purposes, without recognizing the theoretical implications.

Estimation of the Fixed Parameters

The vector y has marginal density $y \sim \mathrm{N}(X\beta, V)$, where $V = \Sigma + ZDZ'$ is specified through the variance component parameters θ. The log-likelihood of the fixed parameters (β, θ) is

$$\ell(\beta,\theta|y) = -\frac{1}{2}\log|V| - \frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta), \tag{1.5}$$

and for fixed θ the estimate $\hat{\beta}$ of β is obtained as the solution of

$$(X'V^{-1}X)\beta = X'V^{-1}y. \tag{1.6}$$

Substituting $\hat{\beta}$ from (1.6) into $\ell(\beta,\theta|y)$ from (1.5) returns the *profile* log-likelihood

$$\begin{aligned}\ell_P(\theta \mid y) &= \ell(\hat{\beta},\theta \mid y)\\ &= -\frac{1}{2}\log|V| - \frac{1}{2}(y - X\hat{\beta})'V^{-1}(y - X\hat{\beta})\end{aligned}$$

of the variance parameter θ. Estimates of the parameters θ specifying V can be found by maximizing $\ell_P(\theta \mid y)$ over θ. These are the ML estimates.

For REML estimation the *restricted* log-likelihood is defined as

$$\ell_R(\theta \mid y) = \ell_P(\theta \mid y) - \frac{1}{2}\log|X'VX|.$$

The REML approach does not base estimates on a maximum likelihood fit of all the information, but instead uses a likelihood function derived from a data set, transformed to remove the irrelevant influences (Dodge, 2003). Restricted maximum likelihood is often preferred to maximum likelihood because REML estimation reduces the bias in the variance component by taking into account the loss of degrees of freedom that results from estimating the fixed effects in $l\beta$. Restricted maximum likelihood also handles high correlations more effectively, and is less sensitive to outliers than maximum likelihood. The problem with REML for model building is that the likelihoods obtained for different fixed effects are not comparable. Hence it is not valid to compare models with different fixed effects using a likelihood ratio test or AIC when REML is used to estimate the model. Therefore models derived using ML must be used instead.

Estimation of the Random Effects

The established approach for estimating the random effects is to use the best linear predictor of b from y, which for a given β equals $GZ'V^{-1}(y - X\beta)$. In practice β is replaced by an estimator such as $\hat{\beta}$ from (1.6) so that $\hat{b} = GZ'V^{-1}(y - X\hat{\beta})$. Pre-multiplying by the appropriate matrices it is straightforward to show that these estimates $\hat{\beta}$ and $\hat{b}$ satisfy the equations in (1.4).

Algorithms for Likelihood Function Optimization

Iterative numerical techniques are used to optimize the log-likelihood function and estimate the covariance parameters θ. The procedure is subject to the constraint that R and D are both positive definite. The most common iterative algorithms for optimizing the likelihood function are the Newton-Raphson method, which is the preferred method, the expectation maximization (EM) algorithm and the Fisher scoring methods.

The EM algorithm, introduced by Dempster et al. (1977), is an iterative technique for maximizing complicated likelihood functions. The algorithm alternates between performing an expectation (E) step and the maximization (M) step. The 'E' step computes the expectation of the log-likelihood evaluated using the current estimate for the variables. In the 'M' step, parameters that maximize the expected log-likelihood, found on the previous 'E' step, are computed. These parameter estimates are then used to determine the distribution of the variables in the next 'E' step. The algorithm alternatives between these two steps until convergence is reached.

The main drawback of the EM algorithm is its slow rate of convergence. Consequently the EM algorithm is rarely used entirely in LME estimation, instead providing an initial set of values that can be passed to other optimization techniques.

The Newton-Raphson (NR) method is the most common, and recommended technique for ML and REML estimation. The NR algorithm minimizes an objective function defines as -2 times the log likelihood for the covariance parameters θ. At every iteration the NR algorithm requires the calculation of a vector of partial derivatives, known as the gradient, and the second derivative matrix with respect to the covariance parameters. This is known as the observed Hessian matrix. Due to the Hessian matrix, the NR algorithm is more time-consuming, but convergence is reached with fewer iterations compared to the EM algorithm. The Fisher scoring algorithm is an variant of the NR algorithm that is more numerically stable and likely to converge, but not recommended to obtain final estimates.

The Extended Likelihood

The desire to have an entirely likelihood-based justification for estimates of random effects, in contrast to Henderson's equation, has motivated Pawitan (2001, page 429) to define the *extended likelihood*. He remarks "In mixed effects modelling the extended likelihood has been called *h-likelihood* (for hierarchical likelihood) by Lee and Nelder (1996), while in smoothing literature it is known as the *penalized likelihood* (e.g. Green and Silverman 1994)." The extended likelihood can be written $L(\beta, \theta, b|y) = p(y|b; \beta, \theta)p(b; \theta)$ and adopting the same distributional assumptions used by **?** yields the log-likelihood function

$$\begin{aligned}\ell_h(\beta, \theta, b|y) \; &= -\frac{1}{2}\left\{\log|\Sigma| + (y - X\beta - Zb)'\Sigma^{-1}(y - X\beta - Zb)\right. \\ &\qquad \left. + \log|D| + b'D^{-1}b\right\}.\end{aligned}$$

Given θ, differentiating with respect to β and b returns Henderson's equations in (1.4).

The LME model as a general linear model

Henderson's equations in (1.4) can be rewritten $(T'W^{-1}T)\delta = T'W^{-1}y_a$ using

$$\delta = \begin{pmatrix} \beta \\ b \end{pmatrix}, \ y_a = \begin{pmatrix} y \\ \psi \end{pmatrix}, \ T = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix}, \ \text{and } W = \begin{pmatrix} \Sigma & 0 \\ 0 & G \end{pmatrix},$$

where **?** describe $\psi = 0$ as quasi-data with mean $\mathrm{E}(\psi) = b$. Their formulation suggests that the joint estimation of the coefficients β and b of the linear mixed effects model can be derived via a classical augmented general linear model $y_a = T\delta + \varepsilon$ where $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) = W$, with *both* β and b appearing as fixed parameters. The usefulness of this reformulation of an LME as a general linear model will be revisited.

1.2 Repeated Measurements in LME models

In many statistical analyzes, the need to determine parameter estimates where multiple measurements are available on each of a set of variables often arises. Further to Lam et al. (1999), Hamlett et al. (2004) performs an analysis of the correlation of replicate measurements, for two variables of interest, using LME models.

Let y_{Aij} and y_{Bij} be the jth repeated observations of the variables of interest A and B taken on the ith subject. The number of repeated measurements for each variable may differ for each individual. Both variables are measured on each time points. Let n_i be the number of observations for each variable, hence $2 \times n_i$ observations in total.

It is assumed that the pair y_{Aij} and y_{Bij} follow a bivariate normal distribution.

$$\begin{pmatrix} y_{Aij} \\ y_{Bij} \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma) \text{ where } \mu = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}$$

The matrix Σ represents the variance component matrix between response variables at a given time point j.

$$\Sigma = \begin{pmatrix} \sigma^2_A & \sigma_{AB} \\ \sigma_{AB} & \sigma^2_B \end{pmatrix}$$

σ^2_A is the variance of variable A, σ^2_B is the variance of variable B and σ_{AB} is the covariance of the two variable. It is assumed that Σ does not depend on a particular time point, and is the same over all time points.

1.2.1 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.

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

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.

1.2.2 Correlation Terms

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

$$D = \begin{pmatrix} \sigma^2_A\rho_A & \sigma_A\sigma_b\rho_{AB}\delta \\ \sigma_A\sigma_b\rho_{AB}\delta & \sigma^2_B\rho_B \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \sigma^2_A(1-\rho_A) & \sigma_{AB}(1-\delta) \\ \sigma_{AB}(1-\delta) & \sigma^2_B(1-\rho_B) \end{pmatrix}.$$

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

Lam et al. (1999) used ML estimation to estimate the true correlation between the variables when the measurements are linked over time. The methodology relies on the assumption that the two variables with repeated measures follow a multivariate normal distribution. The methodology currently does not extend to any more than two cases. The MLE of the correlation takes into account the dependency among repeated measures.

The true correlation ρ_{xy} is repeated measurements can be considered as having two components: between subject and within-subject correlation. The usefulness of estimating repeated measure correlation coefficients is the calculation of between-method and within-method variabilities are produced as by-products.

The Variance Covariance Matrix

The LME model can be written

$$y_i = X_i\beta + Z_ib_i + \epsilon_i,$$

where $\beta = (\beta_0, \beta_1, \beta_2)'$ is a vector of fixed effects, and X_i is a corresponding $2n_i \times 3$ design matrix for the fixed effects. The random effects are expressed in the vector $b = (b_1, b_2)'$, with Z_i the corresponding $2n_i \times 2$ design matrix. The vector ϵ_i is a $2n_i \times 1$ vector of residual terms. Random effects and residuals are assumed to be independent of each other. The variance matrix of $\mathbf{Y}$, denoted $\mathbf{V}$, is an $n \times n$ matrix that can be expressed as follows;

$$\mathbf{V} = \text{Var}(\mathbf{Xb} + \mathbf{Zb} + \mathbf{e}) \tag{1.7}$$

$$\mathbf{V} = \text{Var}(\mathbf{Xb}) + \text{Var}(\mathbf{Zb}) + \text{Var}(\mathbf{e}) \tag{1.8}$$

Var($\mathbf{Xb}$) is known to be zero. The variance of the random effects Var($\mathbf{Zu}$) can be written as $Z\text{Var}(\mathbf{b})Z^T$.

$$\text{var}\begin{pmatrix} b \\ \epsilon \end{pmatrix} = \begin{pmatrix} D & 0 \\ 0 & \Sigma \end{pmatrix}$$

where G and Σ are positive definite matrices parameterized by an unknown variance component parameter vector θ. The variance-covariance matrix for the vector of observations y is given by $V = ZDZ' + \Sigma$. This implies $y \sim (X\beta, V) = (X\beta, ZGZ' + \Sigma)$.

R_i is the variance covariance matrix for the residuals, i.e. the within-item sources of variation between both methods. Computational analysis of linear mixed effects models allow for the explicit analysis of both D and R_i. The above terms can be used to express the variance covariance matrix Ω_i for the responses on item i ,

$$\Omega_i = Z_iGZ_i' + R_i.$$

It is assumed that $b_i \sim N(0, D)$, ϵ_i is a matrix of random errors distributed as $N(0, R_i)$ and that the random effects and residuals are independent of each other. Assumptions made on the structures of D and R_i will be discussed in due course.

The random effects are assumed to be distributed as $b_i \sim \mathcal{N}_2(0, G)$. The between-item variance covariance matrix G is constructed as follows:

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

The distribution of the random effects is described as $b_i \sim N(0, G)$. Similarly random errors are distributed as $\epsilon_i \sim N(0, R_i)$. The random effects and residuals are assumed to be independent. The variance-covariance matrix for the vector of observations y is given by $V = ZGZ' + \Sigma$. This implies $y \sim (X\beta, V) = (X\beta, ZGZ' + \Sigma)$.

By letting $\text{var}(b) = G$ (i.e $\mathbf{b}$ $N(0, \mathbf{G})$), this becomes ZGZ^T. This specifies the covariance due to random effects. The residual covariance matrix $\text{var}(e)$ is denoted as R, ($\mathbf{e}$ $N(0, \mathbf{R})$). Residual are uncorrelated, hence $\mathbf{R}$ is equivalent to $\sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The variance matrix $\mathbf{V}$ can therefore be written as;

$$\mathbf{V} = \mathbf{ZDZ}^T + \mathbf{R} \tag{1.9}$$

1.3 LME Models in Method Comparison Studies

Linear mixed effects (LME) models can facilitate greater understanding of the potential causes of bias and differences in precision between two sets of measurement. Due to computation complexity, linear mixed effects models have not seen widespread use until many well known statistical software applications began facilitating them. Consequently LME approaches have seen increased use as a framework for method comparison studies in recent years (Lai & Shaio, Carstensen and Choudhary as examples)

In part this is due to the increased profile of LME models, and furthermore the availability of capable software. Additionally a great understanding of residual analysis and influence analysis for LME models has been adchieved thanks to authors such as Schabenberger (2004), Christensen et al. (1992), Cook (1986) West et al. (2007), amongst others.

Due to the prevalence of modern statistical software, Carstensen et al. (2008) advocates the adoption of computer based approaches to method comparison studies, allowing the use of LME models that would not have been feasible otherwise. These authors 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'. Rather than using the 'by hand' methods, 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.

Barnhart et al. (2007) sets out three criteria for two methods to be considered in agreement. Firstly that there be no significant bias. Second that there is no difference in the between-subject variabilities, and lastly that there is no significant difference in the within-subject variabilities. Roy further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal.

Varying degrees of importances should be attached to each the three agreement criteria listed by Barnhart et al. (2007). Between-item variance g_i^2 is fundamentally a measure of the variability of the item-wise means, as measured by method i, but it does contain limited information on the precision of that method.

For conventional method comparison problems, both methods measures the same set of items using the same unit of measurement. Convergence to equality of between-item variance inevitable as the number of items n increases. Significantly different estimates for g_1^2 and g_2^2 should not be expected for any practical problem.

Therefore a violation of third criterium (i.e. different between-item variances) criterium is contingent upon, and a possible consequence of, the violation of the other two agreement criteria. However, a violation of the third criterium will not occur in isolation. As noted elsewhere, the matter of inter-method bias can be easily accounted for, once detected. Both between-items and within-items variances must be calculated such that sources of variances are properly assigned, and to compute limits of agreement. However, testing the within-item criterium is the most informative analysis and therefore requires the most attention.

Lai and Shiao (2005) views the uses of linear mixed effects models as an expansion on the Bland-Altman approach, rather than as a replacement. Their focus is to explain lack of agreement by means of additional covariates outside the scope of the traditional method comparison problem, which extends beyond the conventional method comparison study question. The data used for their examples is unavailable for independent use.

1.3.1 Roy's Approach

For the purposes of comparing two methods of measurement, Roy (2009) presents a technique that uses LME modeling. 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 null hypothesis cases.

Two methods of measurement are in complete agreement if the null hypotheses $\mathrm{H}_1\colon \alpha_1 = \alpha_2$ and $\mathrm{H}_2\colon \sigma^2_1 = \sigma^2_2$ and $\mathrm{H}_3\colon d^2_1 = d^2_2$ 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.

Importantly Roy (2009) further proposes a series of three tests on the variance components of an LME model, which allow decisions on the second and third of Roy's criteria. 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,

$$A = \left(\begin{array}{cc} a_{11} & a_{12} \\ a_{21} & a_{22} \end{array} \right).$$

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

The first model acts as an alternative hypothesis to be compared against each of three other models, acting as null hypothesis models, successively. The models are compared using the likelihood ratio test. Likelihood ratio tests are a class of tests based on the comparison of the values of the likelihood functions of two candidate models. LRTs can be used to test hypotheses about covariance parameters or fixed effects parameters in the context of LMEs. The test statistic for the likelihood ratio test is the difference of the log-likelihood functions, multiplied by -2.

The probability distribution of the test statistic is approximated by the χ^2 distribution with $(\nu_1 - \nu_2)$ degrees of freedom, where ν_1 and ν_2 are the degrees of freedom of models 1 and 2 respectively. Each of these three test shall be examined in more detail shortly.

1.3.2 Replicate Measurements in Roy's paper

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.

1.3.3 Test For Inter-Method Bias

Firstly, a practitioner would investigate whether a significant inter-method bias is present between the methods. This bias is specified as a fixed effect in the LME model. For a practitioner who has a reasonable level of competency in statistical software and undergraduate statistics (in particular simple linear regression model) this is a straight-forward procedure. The presence of an inter-method bias is the source of disagreement between two methods of measurement that is most easily identified. As the first in a series of hypothesis tests, Roy (2009) presents a formal test for inter-method bias. With the null and alternative hypothesis denoted H_1 and K_1 respectively, this test is formulated as

$$\mathrm{H}_1 : \mu_1 = \mu_2,$$

$$\mathrm{K}_1 : \mu_1 \neq \mu_2.$$

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

1.3.4 Roy's Hypothesis Tests For Variability

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.

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 (**?**).

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.

$$\mathrm{H}_3 : \sigma^2_1 = \sigma^2_2$$
$$\mathrm{K}_3 : \sigma^2_1 \neq \sigma^2_2$$

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

$$\mathrm{H}_4 : \omega^2_1 = \omega^2_2$$
$$\mathrm{K}_4 : \omega^2_1 \neq \omega^2_2$$

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

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.

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 procedure uses a linear mixed effects regression fit using a combination of symmetric and compound symmetry (CS) correlation structure the variance covariance matrices.

1.3.5 Model Specification for Roy's Hypotheses Tests

Response for ith subject can be written as

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

- β_1 and β_2 are fixed effects corresponding to both methods. (β_0 is the intercept.)
- b_{1i} and b_{2i} are random effects corresponding to both methods.

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 $Y_i = (y_{1i1}, y_{2i1}, y_{1i2}, \ldots, y_{mir}, \ldots, y_{1in_i}, y_{2in_i})'$.

1.3.6 Specifying the Models

Roy proposes a series of three tests on the variance components of an LME model. For these tests, four candidate models are constructed, each differing by constraints applied to the variance covariance matrices. In addition to computing the inter-method bias, three significance tests are carried out on the respective formulations to make a judgement on whether or not two methods are in agreement.

The difference in the models are specifically in how the the G 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.

1.3.7 Variability Tests

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

$$H_0: \; g_1 = g_2$$
$$H_A: \; g_1 \neq g_2$$

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

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 $\hat{G}$ and $\hat{\Sigma}$. The null model is constructed a symmetric form for $\hat{\Sigma}$ while the alternative model uses a compound symmetry form. This time $\hat{G}$ 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 also integrates H$_2$ and H$_3$ into a single testable hypothesis H$_4$: $\omega^2_1 = \omega^2_2$, where $\omega^2_m = \sigma^2_m + d^2_m$ 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.

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 $\hat{\Sigma}$.

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \tag{1.10}$$

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

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

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 $\hat{G}$ and Λ while the alternative model uses a compound symmetry form for both.

1.4 Interpreting the Correlation Coefficient

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

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.

Roy (2009) remarks that 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.

The approach proposed by Roy (2009) 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.

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.

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

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.

1.4.1 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 approach incorporates the use of correlation.

In addition to the variability tests, Roy advises that it is preferable that a correlation of greater than 0.82 exist for two methods to be considered interchangeable. However if two methods fulfil all the other conditions for agreement, failure to comply with this one can be overlooked. Indeed Roy demonstrates that placing undue importance to it can lead to incorrect conclusions.

Roy (2009) remarks that current computer implementations only gives overall correlation coefficients, but not their variances. Consequently it is not possible to carry out inferences based on all overall correlation coefficients.

1.4.2 Formal Testing for Covariances

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

$$\begin{pmatrix} \sigma^2_x & \sigma_{xy} \\ \sigma_{xy} & \sigma^2_y \end{pmatrix}$$

σ^2_x and σ^2_y 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.

$$\begin{pmatrix} \sigma^2_x & \sigma_x y \\ \sigma_x y & \sigma^2_y \end{pmatrix} vs \begin{pmatrix} \sigma^2_x & 0 \\ 0 & \sigma^2_y \end{pmatrix}$$

As it is pertinent to the difference between the two described approaches, the facilitation of a formal test would be useful. Extending the approach proposed by ARoy2009, the test for overall covariance can be formulated:

$$\mathrm{H}_5 : \sigma_{12} = 0$$

$$\mathrm{K}_5 : \sigma_{12} \neq 0$$

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.

1.5 Extension of Roy's Technique

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

An implementation of Roy's approach, 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.

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

Roy's approach 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.

1.6 Conclusion

Carstensen et al. (2008) and Roy (2009) highlight the need for method comparison approaches suitable for use in the presence of replicate measurements. Roy (2009) presents a comprehensive system for assessing the agreement of two methods, for replicate measurements. This approach has the added benefit of overcoming the problems of unbalanced data and unequal numbers of replicates. Implementation of the approach, and interpretation of the results, can be made easy for practitioners that have only basic statistical training. Furthermore, it can be shown that widely used procedures, such as the limits of agreement, can be incorporated into Roy's approach.

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., 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).

Christensen, R., L. M. Pearson, and W. Johnson (1992). Case-deletion diagnostics for mixed models. *Technometrics 34*(1), 38–45.

Cook, R. (1986). Assessment of local influence. *Journal of the Royal Statistical Society. Series B (Methodological) 48*(2), 133–169.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the em algorithme. *ournal of the Royal Statistical Society. Series B 39*(1), 1–38.

Dodge, Y. (2003). *The Oxford Dictionary of Statistical Terms.* Oxford University Press.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics 3*(1), 1–21.

Fisher, R. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh 2*, 399–433.

Green, P. J. and B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach.* Chapman & Hall Ltd.

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.

Hartley, H. and J. Rao (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika 54*(1/2), 93–108.

Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics 9*(2), 226–252.

Henderson, C., O. Kempthorne, S. Searle, and C. von Krosigk (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192–218.

Henderson, C., O. Kempthorne, S. Searle, and C. von Krosigk (1963). The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192–218.

Henderson, C., O. Kempthorne, S. Searle, and C. von Krosigk (1973). The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192–218.

Henderson, C., O. Kempthorne, S. Searle, and C. von Krosigk (1984). The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192–218.

Henderson, C. R. (1973). Sire evaluation and genetic trends. In *Proceedings of the Animal Breeding and genetics Symposium in Honor of Dr Jay L. Lush*, pp. 10–41. Champaign, Illinois: Americian Society of Animal Science and American Dairy Science Association.

Kelly, G. E. (1985). Use of the structural equations model in assessing the reliability of a new measurement technique. *Applied Statistics*, 258–263.

Lai, D. and S.-Y. P. K. Shiao (2005). Comparing two clinical measurements: a linear mixed model approach. *Journal of Applied Statistics 32*(8), 855–860.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics 38*(4), 963–974.

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

Lee, Y., J. Nelder, and Y. Pawitan (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood.* Chapman and Hall CRC.

Lee, Y. and J. A. Nelder (1996). Hierarchical generalized linear models (Disc: P656-678). *Journal of the Royal Statistical Society, Series B: Methodological 58*, 619–656.

Patterson, H. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika 58*(3), 545–554.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford University Press.

Pinheiro, J. and D. Bates (1994). *Mixed Effects Models in S and S plus* (2nd ed.). Reading, Massachusetts: Springer.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects (Disc: P32-51). *Statistical Science 6*, 15–32.

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.

Schabenberger, O. (2004). Mixed model influence diagnostics. 189-29.

Singer, J. D. (1998). Using sas proc mixed to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics 24*(4), 323–355.

Tippett, L. (1931). *The Methods of Statistics* (1st ed.). London: Williams and Norgate.

West, B., K. Welch, and A. Galecki (2007). *Linear Mixed Models: a Practical Guide Using Statistical Software*. Chapman and Hall CRC.