Contents

Chapter 1

Linear Mixed effects Models

1.1 Linear Mixed effects Models

A linear mixed effects (LME) model is a statistical model containing both fixed effects and random effects (random effects are 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 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. Henderson (1950) devised a framework 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 methodology is 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.

1.1.1 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 $\hat{b}$. Henderson (1950) made the (ad-hoc) distributional assumptions $y|b \sim \mathrm{N}(X\beta + Zb, \Sigma)$ and $b \sim \mathrm{N}(0, D)$, and proceeded to maximize the joint density of y and b

$$\left| \begin{array}{cc} D & 0 \\ 0 & \Sigma \end{array} \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix}' \begin{pmatrix} D & 0 \\ 0 & \Sigma \end{pmatrix}^{-1} \begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix} \right\}, \quad (1.1)$$

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

$$(y - X\beta - Zb)'\Sigma^{-1}(y - X\beta - Zb) + b'D^{-1}b. \quad (1.2)$$

This leads to the mixed model equations

$$\begin{pmatrix} X'\Sigma^{-1}X & X'\Sigma^{-1}Z \\ Z'\Sigma^{-1}X & X'\Sigma^{-1}X + D^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X'\Sigma^{-1}y \\ Z'\Sigma^{-1}y \end{pmatrix}. \quad (1.3)$$

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 D and Σ are known.

Robinson (1991) points out that although Henderson (1950) initially referred to the estimates $\hat{\beta}$ and $\hat{b}$ from (1.3) 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 (1.2) is a joint density rather than a likelihood function. Lee et al. (2006a) remarks that it is clear that Henderson used joint estimation for computational purposes, without recognizing the theoretical implications.

Hartley and Rao (1967) demonstrated that unique estimates of the variance components could be obtained using maximum likelihood methods. The maximum likelihood procedure of Hartley and Rao yields simultaneous estimates for both the fixed effects and the random effect, by maximising the likelihood of $\boldsymbol{y}$ with respect to each element of $\boldsymbol{\beta}$ and $\boldsymbol{b}$.

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.

1.1.2 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. Using Laird-Ware form, the LME model is commonly described in matrix form,

$$Y = X\beta + Zb + \epsilon \tag{1.4}$$

$\mathbf{Y}$ 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. $\mathbf{e}$ is the vector of residuals with dimension $n \times 1$. The random effects are contained in the $q \times 1$ vector $\mathbf{b}$.

The Variance Covariance Matrix

$\mathbf{V}$, the variance matrix of $\mathbf{Y}$, can be expressed as follows;

$$\mathbf{V} = \text{Var}(\mathbf{Xb} + \mathbf{Zb} + \mathbf{e}) \tag{1.5}$$

$$\mathbf{V} = \text{Var}(\mathbf{Xb}) + \text{Var}(\mathbf{Zb}) + \text{Var}(\mathbf{e}) \tag{1.6}$$

$\text{Var}(\mathbf{Xb})$ is known to be zero. The variance of the random effects $\text{Var}(\mathbf{Zu})$ 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 D 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 $y \sim (X\beta, V) = (X\beta, ZDZ' + \Sigma)$. It is worth noting that V is an $n \times n$ matrix, as the dimensionality becomes relevant later on. The notation provided here is generic, and will be adapted to accord with complex formulations that will be encountered in due course.

The variance-covariance matrix for the vector of observations y is given by $V = ZDZ' + \Sigma$. This implies $y \sim (X\beta, V) = (X\beta, ZDZ' + \Sigma)$.

By letting $\text{var}(b) = D$ (i.e $\mathbf{b}$ $N(0, \mathbf{D})$), this becomes ZDZ^T. This specifies the covariance due to random effects. The residual covariance matrix $var(e)$ 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;

$$\mathbf{V} = \mathbf{ZDZ}^T + \mathbf{R} \tag{1.7}$$

Decomposition of the response covariance matrix

The variance covariance structure, $\boldsymbol{\Omega_i}$, can be re-expressed in the following form,

$$\mathrm{Cov}(\mathrm{y}_i) = \boldsymbol{\Omega_i} = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' + \boldsymbol{R_i}.$$

$\boldsymbol{R_i}$ can be shown to be the Kronecker product of a correlation matrix $\boldsymbol{V}$ and $\boldsymbol{\Lambda}$. The correlation matrix $\boldsymbol{V}$ of the repeated measures on a given response variable is assumed to be the same for all response variables. Both Hamlett et al. (2004) and Lam et al. (1999) use the identity matrix, with dimensions $n_i \times n_i$ as the formulation for $\boldsymbol{V}$. Roy (2009) remarks that, with repeated measures, the response for each subject is correlated for each variable, and that such correlation must be taken into account in order to produce a valid inference on correlation estimates. Roy (2006) proposes various correlation structures may be assumed for repeated measure correlations, such as the compound symmetry and autoregressive structures, as alternative to the identity matrix.

However, for the purposes of method comparison studies, the necessary estimates are currently only determinable when the identity matrix is specified, and the results in Roy (2009) indicate its use.

The distribution of the random effects is described as $\boldsymbol{b}_i \sim N(0, \boldsymbol{D})$. Similarly random errors are distributed as $\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{R}_i)$. The random effects and residuals are assumed to be independent.

It is important to note that no special assumptions about the structure of $\boldsymbol{D}$ are made. An example of such an assumption would be that $\boldsymbol{D}$ is the product of a scalar value and the identity matrix.

The matrix of random errors $\boldsymbol{\epsilon}_i$ is distributed as $\mathcal{N}_2(0, \boldsymbol{R}_i)$. Hamlett et al. (2004) shows that the variance covariance matrix for the residuals(i.e. the within-item sources of variation between both methods) can be expressed as the Kroneckor product of an $n_i \times n_i$ identity matrix and the partial within-item variance covariance matrix $\boldsymbol{\Sigma}$, i.e.

$\boldsymbol{R}_i = \boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}$.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

where σ_1^2 and σ_2^2 are the within-subject variances of the respective methods, and σ_{12} is the within-item covariance between the two methods.

The within-item variance covariance matrix $\boldsymbol{\Sigma}$ is assumed to be the same for all replications. Computational analysis of linear mixed effects models allow for the explicit analysis of both $\boldsymbol{D}$ and $\boldsymbol{R_i}$.

The partial within-item variance-covariance matrix of two methods at any replicate is denoted $\boldsymbol{\Sigma}$, where σ_1^2 and σ_2^2 are the within-subject variances of the respective methods, and σ_{12} is the within-item covariance between the two methods.

For the response vector described, Hamlett et al. (2004) presents a detailed covariance matrix. A brief summary shall be presented here only. The overall variance matrix is a 6×6 matrix composed of two types of 2×2 blocks. Each block represents one separate time of measurement.

$$\boldsymbol{\Omega}_i = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{D} & \boldsymbol{D} \\ \boldsymbol{D} & \boldsymbol{\Sigma} & \boldsymbol{D} \\ \boldsymbol{D} & \boldsymbol{D} & \boldsymbol{\Sigma} \end{pmatrix}$$

The diagonal blocks are Σ, as described previously. The 2×2 block diagonal matrix in $\boldsymbol{\Omega}$ gives $\boldsymbol{\Sigma}$. $\boldsymbol{\Sigma}$ is the sum of the between-subject variability $\boldsymbol{D}$ and the within subject variability $\boldsymbol{\Lambda}$ (in Hamletts' notation). $\boldsymbol{\Omega_i}$ can be expressed as

$$\boldsymbol{\Omega_i} = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' + (\boldsymbol{I}_{\boldsymbol{n_i}} \otimes \boldsymbol{\Lambda}).$$

Hamlett et al. (2004) shows that the variance covariance matrix for the residuals(i.e. the within-item sources of variation between both methods) can be expressed as the Kroneckor product of an $n_i \times n_i$ identity matrix and the partial within-item variance covariance matrix $\boldsymbol{\Sigma}$, i.e. $\boldsymbol{R}_i = \boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}$.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

where σ_1^2 and σ_2^2 are the within-subject variances of the respective methods, and σ_{12} is the within-item covariance between the two methods.

The within-item variance covariance matrix $\boldsymbol{\Sigma}$ is assumed to be the same for all replications. Again it is important to note that no special assumptions are made about the structure of the matrix. Computational analysis of linear mixed effects models allow for the explicit analysis of both $\boldsymbol{D}$ and $\boldsymbol{R_i}$.

$$\boldsymbol{R}_i = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & 0 & 0 & \dots & \dots & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & 0 & 0 & \dots & \dots & 0 & 0 \\ 0 & 0 & \sigma_1^2 & \sigma_{12} & \dots & \dots & 0 & 0 \\ 0 & 0 & \sigma_{12} & \sigma_2^2 & \dots & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \dots & \sigma_1^2 & \sigma_{12} \\ 0 & 0 & 0 & 0 & \dots & \dots & \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

Hamlett et al. (2004) shows that $\boldsymbol{R}_i$ can be expressed as $\boldsymbol{R}_i = \boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}$.

The covariance matrix has the same structure for all items, except for dimension, which depends on the number of replicates.

Formulation of the response vector

Information of individual i is recorded in a response vector $\boldsymbol{y}_i$. 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 $\boldsymbol{y_i}$ 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, $\boldsymbol{y}_i$ is a 6×1 random vector describing the ith subject.

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

The response vector $\boldsymbol{y_i}$ can be formulated as an LME model according to Laird-Ware form.

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

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

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.

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

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

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix}$$

σ_A^2 is the variance of variable A, σ_B^2 is the variance of variable B and σ_{AB} is the covariance of the two variable. It is assumed that $\boldsymbol{\Sigma}$ does not depend on a particular time point, and is the same over all time points.

Correlation terms

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

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

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

ρ_A describe the correlations of measurements made by the method A at different times. Similarly ρ_B describe the correlation of measurements made by the method B at different times. Correlations among repeated measures within the same method are known as intra-class correlation coefficients. 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.

There is a substantial difference in the number of fixed parameters used by the respective models; the model in (1.8) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items N, whereas the model in (**??**) requires $N + 2$ fixed effects.

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

1.1.3 Likelihood and estimation

Likelihood is the hypothetical probability that an event that has already occurred would yield a specific outcome. Likelihood differs from probability in that probability refers to future occurrences, while likelihood refers to past known outcomes.

Likelihood functions provide the basis for two important statistical concepts that shall be further referred to; the likelihood ratio test and the Akaike information criterion. The likelihood function, $L(\theta)$, is a fundamental concept in statistical inference. It indicates how likely a particular population is to produce an observed sample. The set of values that maximize the likelihood function are considered to be optimal, and are used as the estimates of the parameters. For computational ease, it is common to use the logarithm of the likelihood function, known simply as the log-likelihood ($\ell(\theta)$).

Assuming a statistical model $f_\theta(y)$ parameterized by a fixed and unknown set of parameters θ, the likelihood $L(\theta)$ is the probability of the observed data y considered as a function of θ (Lee et al., 2006a).

Likelihood-based tools

The maximum likelihood estimates (MLEs) of the parameters are the values of the arguments that maximize the likelihood function. Maximum likelihood and restricted maximum likelihood have become the most common strategies for estimating the variance component parameter θ. Maximum likelihood (ML) estimation is a well known method of obtaining estimates of unknown parameters by optimizing a likelihood function. To obtain ML estimate the likelihood is constructed as a function of the parameters in the specified LME model.

The likelihood function is constructed as a function of the parameters in the specified model. Models fitted by ML estimation can be compared using the likelihood ratio test. However ML is known to underestimate variance components for finite samples (Demidenko, 2004).

1.1.4 Algorithms

Maximum likelihood estimation is a method of obtaining estimates of unknown parameters by optimizing a likelihood function. The ML parameter estimates are the values of the argument that maximise the likelihood function, i.e. the estimates that make the observed values of the dependent variable most likely, given the distributional assumptions

The most common iterative algorithms used for the optimization problem in the context of LMEs are the EM algoritm, fisher scoring algorithm and NR algorithm, which West et al. (2007) commends as the preferred method. Parameter of the mixed model can be estimated using either ML or REML, while the AIC and the BIC can be used as measures of "goodness of fit" for particular models, where smaller values are

considered preferable.

Restricted Likelihood Estimation

Restricted maximum likelihood (REML) is an alternative methods of computing parameter estimated, developed by Paterson and Thompson (1971) and Harville (1977) to provide unbiased estimates of variance and covariance parameters. The REML approach is a variant of maximum likelihood estimation which 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).

REML obtains estimates of the fixed effects using non-likelihoodlike methods, such as ordinary least squares or generalized least squares, and then using these estimates it maximizes the likelihood of the residuals (subtracting off the fixed effects) to obtain estimates of the variance parameters. In most software packages REML is the default algorithm used to compute coefficients for the predictor variables.

In contrast to the earlier maximum likelihood estimation, REML can produce unbiased estimates of variance and covariance parameters. REML estimation reduces the bias in the variance component, and also handles high correlations more effectively, and is less sensitive to outliers than ML.

The variance components in the LME model may be estimated by ML or REML. Maximum Likelihood estimates do not take into account the estimation of fixed effects and so are biased downwards. REML estimates accounts for the presence of these nuisance parameters by maximising the linearly independent error contrasts to obtain more unbiased estimates.

McCullough and Searle (2001) describes two important outcomes of using REML. Firstly variance components can be estimated without being affected by fixed effects. Secondly in estimating variance components with REML, degrees of freedom for the fixed effects can be taken into account implicitly, whereas with ML they are not.

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 $\boldsymbol{\beta}$. 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.

1.1.5 Estimation of Fixed and Random Effects

Potentially it may be impossible to compute unique BLUE estimates for all the fixed factors in a model. This may be due to linear dependence in the model matrix $\mathbf{X}$.

Estimation of random effects for LME models in the NLME package is accomplished through use of both EM (Expectation-Maximization) algorithms and Newton-Raphson algorithms.

- EM iterations bring estimates of the parameters into the region of the optimum very quickly, but convergence to the optimum is slow when near the optimum.
- Newton-Raphson iterations are computationally intensive and can be unstable when far from the optimum. However, close to the optimum they converge quickly.
- The LME function implements a hybrid approach, using 25 EM iterations to quickly get near the optimum, then switching to Newton-Raphson iterations to quickly converge to the optimum.
- If convergence problems occur, the "control argument in LME can be used to change the way the model arrives at the optimum.

1.2 Introduction

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. 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.

While the method comparison problem is conventionally poised in the context of two methods of measurements, LME models allow for a straightforward analysis whereby several methods of measurement can be measured simulataneously. However simple models only can only indicate agreement of lack thereof, and the presence of inter-method bias. To consider more complex questions, more complex LME models are required. Useful approaches will be introduced in a later section.

1.2.1 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) describes the sources of disagreement in a method comparison study problem as differing population means, different between-subject variances, different within-subject variances between two methods and poor correlation between measurements of two methods. Further to this, Roy (2009) states 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 (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.

Lai and Shiao (2005) views the uses of linear mixed effects models as an expansion on the Bland-Altman methodology, 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. Therefore, for the sake of consistency, a data set will be simulated based on the Blood Data that will allow for extra variables, and an exploration shall be provided in the appendices.

1.2.2 Fitting LME Models to Method Comparison Data

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.

1.2.3 Roy's Approach

For the purposes of comparing two methods of measurement, Roy (2009) presents a methodology utilizing linear mixed effects model. This methodology provides for the formal testing of inter-method bias, between-subject variability and within-subject variability of two methods.

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.

Roy (2009) proposes the use of LME models to perform a test on two methods of agreement to comparing the agreement between two methods of measurement, where replicate measurements on items (often individuals) by both methods are available, determining whether they can be used interchangeably. The methodology proposed by

Roy (2009) is largely based on Hamlett et al. (2004), which in turn follows on from Lam et al. (1999).

This approach uses a Kronecker product covariance structure with doubly multivariate setup to assess the agreement, and is designed such that the data may be unbalanced and with unequal numbers of replications for each subject (Roy, 2009).

Roy (2009) proposes a novel method using the LME model with Kronecker product covariance structure in a doubly multivariate set-up to assess the agreement between a new method and an established method with unbalanced data and with unequal replications for different subjects.

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 an approach based on linear mixed effects (LME) models for the purpose of comparing the agreement between two methods of measurement, where replicate measurements on items (often individuals) by both methods are available. Three tests of hypothesis appropriate are provided for evaluating the agreement between the two methods of measurement under this sampling scheme.

The well-known "Limits of Agreement", as developed by Bland and Altman (1986) are not referred to directly, but are easily computable using the framework proposed by Roy (2009). Further discussion will be provided in due course.

1.2.4 Replicate measurements in Roy's paper

Measurements taken in quick succession by the same observer using the same instrument on the same subject can be considered true replicates. Roy (2009) notes that some measurements may not be 'true' replicates.

Roy's methodology assumes the use of "true replicates". However data may 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) takes its definition of replicate measurement: two or more measurements on the same item taken under identical conditions. Roy also assumes linked measurements, but it is can be used for the non-linked case.

1.2.5 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.

Using Roy's method, 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.

Roy's methodology requires the construction of four candidate models.

The difference in the models are specifically in how the the D and Σ matrices are constructed, using either an unstructured form or a compound symmetry form. The first model is compared against each of three other models successively.

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 and Romano, 2006).

Model 1

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.

Roy (2009) considers four independent hypothesis tests. 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.

- Testing of hypotheses of differences between the means of two methods
- Testing of hypotheses in between subject variabilities in two methods,
- Testing of hypotheses of differences in within-subject variability of the two methods,
- Testing of hypotheses in differences in overall variability of the two methods.

These tests are the pairwise comparison of candidate models, one formulated without constraints, the other with a constraint.

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

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. The first test allows of the comparison the begin-subject variability of two methods. 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.

1.2.6 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 $\boldsymbol{y_i} = (y_{1i1}, y_{2i1}, y_{1i2}, \ldots, y_{mir}, \ldots, y_{1in_i}, y_{2in_i})'$. The LME model can be written

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

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ is a vector of fixed effects, and $\boldsymbol{X}_i$ is a corresponding $2n_i \times 3$ design matrix for the fixed effects. The random effects are expressed in the vector $\boldsymbol{b} = (b_1, b_2)'$, with $\boldsymbol{Z}_i$ the corresponding $2n_i \times 2$ design matrix. The vector $\boldsymbol{\epsilon}_i$ is a $2n_i \times 1$ vector of residual terms. Random effects and residuals are assumed to be independent of each other.

$\boldsymbol{R}_i$ 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 $\boldsymbol{D}$ and $\boldsymbol{R_i}$. The above terms can be used to express the variance covariance matrix $\boldsymbol{\Omega}_i$ for the responses on item i ,

$$\boldsymbol{\Omega}_i = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' + \boldsymbol{R}_i.$$

It is assumed that $\boldsymbol{b}_i \sim N(0, \boldsymbol{D})$, $\boldsymbol{\epsilon}_i$ is a matrix of random errors distributed as $N(0, \boldsymbol{R}_i)$ and that the random effects and residuals are independent of each other. Assumptions made on the structures of $\boldsymbol{D}$ and $\boldsymbol{R}_i$ will be discussed in due course.

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

$$\boldsymbol{D} = \begin{pmatrix} d_1^2 & d_{12} \\ d_{12} & d_2^2 \end{pmatrix}$$

$$\boldsymbol{D} = \text{Var} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{pmatrix} d_1^2 & d_{12} \\ d_{12} & d_2^2 \end{pmatrix}$$

1.2.7 LME Model Specification

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

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \tag{1.8}$$

Here β_0 and β_m are fixed-effect terms representing, respectively, a model intercept and an overall effect for method m.

The b_{1i} and b_{2i} terms represent random effect parameters corresponding to the two methods, having $\text{E}(b_{mi}) = 0$ with $\text{Var}(b_{mi}) = g_m^2$ and $\text{Cov}(b_{mi}, b_{m'i}) = d_{12}$. The random error term for each response is denoted ϵ_{mir} having $\text{E}(\epsilon_{mir}) = 0$, $\text{Var}(\epsilon_{mir}) = \sigma_m^2$, $\text{Cov}(b_{mir}, b_{m'ir}) = \sigma_{12}$, $\text{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$ and $\text{Cov}(\epsilon_{mir}, \epsilon_{m'ir'}) = 0$.

When two methods of measurement are in agreement, there is no significant differences between β_1 and β_2, g_1^2 andg_2^2, and σ_1^2 andσ_2^2. 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 $\text{Var}(b_{mi}) = g_m^2$ and $\text{Cov}(b_{1i}, b_{2i}) = d_{12}$.

The random error term for each response is denoted ϵ_{mir} having $\text{E}(\epsilon_{mir}) = 0$, $\text{Var}(\epsilon_{mir}) = \sigma_m^2$, $\text{Cov}(\epsilon_{1ir}, \epsilon_{2ir}) = \sigma_{12}$, $\text{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$ and $\text{Cov}(\epsilon_{1ir}, \epsilon_{2ir'}) = 0$. Two methods of measurement are in complete agreement if the null hypotheses $\text{H}_1 : \alpha_1 = \alpha_2$

and $H_2 : \sigma_1^2 = \sigma_2^2$ and $H_3 : g_1^2 = g_2^2$ hold simultaneously. Roy (2009) uses a Bonferroni correction to control the familywise error rate for tests of $\{H_1, H_2, H_3\}$ and account for difficulties arising due to multiple testing.

1.2.8 Agreement Criteria

Roy's method considers two methods to be in agreement if three: no significant bias, i.e. the difference between the two mean readings is not "statistically significant", high overall correlation coefficient, the agreement between the two methods by testing their repeatability coefficients. Two methods of measurement can be said to be in agreement if there is no significant difference between in three key respects.

Roy additionally uses the overall correlation coefficient to provide extra information about the comparison, with a minimum of 0.82 being required. Firstly, there is no inter-method bias between the two methods, i.e. there is no persistent tendency for one method to give higher values than the other. Secondly, both methods of measurement have the same within-subject variability. In such a case the variance of the replicate measurements would consistent for both methods. Lastly, the methods have equal between-subject variability. Put simply, for the mean measurements for each case, the variances of the mean measurements from both methods are equal. Lack of agreement can arise if there is a disagreement in overall variabilities. This may be due to due to the disagreement in either between-item variabilities or within-item variabilities, or both. Roy (2009) allows for a formal test of each.

Further to this, Roy (2009) demonstrates an suite of tests that can be used to determine how well two methods of measurement, in the presence of repeated measures, agree with each other.

- No Significant inter-method bias

- No difference in the between-subject variabilities of the two methods

- No difference in the within-subject variabilities of the two methods

The formulation presented above usefully facilitates a series of significance tests that advise as to how well the two methods agree. These tests are as follows:

- A formal test for the equality of between-item variances,
- A formal test for the equality of within-item variances,
- A formal test for the equality of overall variances.

These tests are complemented by the ability to consider the inter-method bias and the overall correlation coefficient. 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. Other important aspects of the method comparison study are consequent. The limits of agreement are computed using the results of the reference model.

Differences in between-subject variabilities of the two methods arise when one method is yielding average response levels for individuals than are more variable than the average response levels for the same sample of individuals taken by the other method. Differences in within-subject variabilities of the two methods arise when one method is yielding responses for an individual than are more variable than the responses for this same individual taken by the other method. The two methods of measurement can be considered to agree, and subsequently can be used interchangeably, if all three null hypotheses are true.

1.2.9 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 R 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.

1.2.10 Roy's hypothesis tests : Roy's variability tests

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 usefully facilitates a series of significance tests that assess if and where such differences arise. Roy allows for a formal test of each. These tests are comprised of a formal test for the equality of between-item variances.

$$\mathrm{H}_3 : \sigma_1^2 = \sigma_2^2$$

$$\mathrm{K}_3 : \sigma_1^2 \neq \sigma_2^2$$

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

$$\mathrm{H}_4 : \omega_1^2 = \omega_2^2$$
$$\mathrm{K}_4 : \omega_1^2 \neq \omega_2^2$$

These tests are complemented by the ability to consider the inter-method bias and the overall correlation coefficient. 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.

1.2.11 Variance Covariance Matrices

Under Roy's model, random effects are defined using a bivariate normal distribution. Consequently, the variance-covariance structures can be described using 2×2 matrices. A discussion of the various structures a variance-covariance matrix can be specified under is required before progressing. The following structures are relevant: the identity structure, the compound symmetric structure and the symmetric structure.

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,

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

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

The identity structure is simply an abstraction of the identity matrix. The compound symmetric structure and symmetric structure can be described with reference to the following matrix (here in the context of the overall covariance Block-$\boldsymbol{\Omega}_i$, but equally applicable to the component variabilities $\boldsymbol{D}$ and $\boldsymbol{\Sigma}$);

$$\begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix}$$

Symmetric structure requires the equality of all the diagonal terms, hence $\omega_1^2 = \omega_2^2$. Conversely compound symmetry make no such constraint on the diagonal elements. Under the identity structure, $\omega_{12} = 0$. A comparison of a model fitted using symmetric structure with that of a model fitted using the compound symmetric structure is equivalent to a test of the equality of variance.

Independence

As though analyzed using between subjects analysis.

$$\begin{pmatrix} \psi^2 & 0 & 0 \\ 0 & \psi^2 & 0 \\ 0 & 0 & \psi^2 \end{pmatrix}$$

Compound Symmetry

Assumes that the variance-covariance structure has a single variance (represented by ψ^2) for all 3 of the time points and a single covariance (represented by ψ_{ij}) for each of the pairs of trials.

$$\begin{pmatrix} \psi^2 & \psi_{12} & \psi_{13} \\ \psi_{21} & \psi^2 & \psi_{23} \\ \psi_{31} & \psi_{32} & \psi^2 \end{pmatrix}$$

1.2.12 Variability Tests

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

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.

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 $\mathrm{H}_2 \colon \sigma_1^2 = \sigma_2^2$ 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.

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

Variability test 1

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

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

This test is facilitated by constructing a model specifying a symmetric form for D (i.e. the alternative model) and comparing it with a model that has compound symmetric

form for D (i.e. the null model). For this test $\hat{\mathbf{\Lambda}}$ 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 $\hat{\boldsymbol{D}}$ and $\hat{\mathbf{\Lambda}}$. The null model is constructed a symmetric form for $\hat{\mathbf{\Lambda}}$ while the alternative model uses a compound symmetry form. This time $\hat{\boldsymbol{D}}$ has a symmetric form for both models, and will be the same for both.

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

Variability test 3

Roy also integrates H_2 and H_3 into a single testable hypothesis $\mathrm{H}_4 \colon \omega_1^2 = \omega_2^2$, where $\omega_m^2 = \sigma_m^2 + g_m^2$ represent the overall variability of method m.

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 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 $\hat{\boldsymbol{D}}$ and $\hat{\boldsymbol{\Lambda}}$ while the alternative model uses a compound symmetry form for both.

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 $\hat{\Sigma}$.

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \tag{1.9}$$

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

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

1.2.13 Computing Limits of Agreement

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 - $\boldsymbol{\Omega}_i$ matrix. Lack of agreement can arise if there is a disagreement in overall variabilities.

1.2.14 Formal testing for covariances (Off-Diagonal Components in Roy's Model)

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.

$$\begin{pmatrix} \sigma_x^2 & \sigma_x y \\ \sigma_x y & \sigma_y^2 \end{pmatrix} vs \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$$

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 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.

1.2.15 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.

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.

1.2.16 Extension of Roy's methodology

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

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

every increase in n.

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

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

1.2.17 Roy's methodology for single 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.

1.3 Likelihood Ratio Tests

Likelihood ratio tests (LRTs) are a family of tests used to compare the value of likelihood functions for two models, whose respective formulations define a hypothesis to be tested (i.e. the nested and reference model). LRTs can be used to test hypotheses

about covariance parameters or fixed effects parameters in the context of LMEs. Each of these three test shall be examined in more detail shortly.

The test statistic for each of the three hypothesis tests is the difference of the $M2LL$ for each pair of models. 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. If the $p-$value in each of the respective tests exceed as significance level chosen by the analyst, then the null model must be rejected. The significance of the likelihood ratio test can be found by comparing the likelihood ratio to the χ^2 distribution, with the appropriate degrees of freedom.

$$\nu = [\text{ LRT df under } H_0 \text{ model}] - [\text{ LRT df under } H_A \text{ model}] \tag{1.10}$$

L= - 2ln is approximately distributed as 2 under H_0 for large sample size and under the normality assumption.

$$-2 \ln \Lambda_d = [M2LL \text{ under } H_0 \text{ model}] - [M2LL \text{ under } H_A \text{ model}] \tag{1.11}$$

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

Such a test can also be used for models fitted using REML, but only if both models have been fitted by REML, and if the fixed effects specification is the same for both models.

Each of these three test shall be examined in more detail shortly. The power of the likelihood ratio test may depends on specific sample size and the specific number of replications, and Roy (2009) proposes simulation studies to examine this further.

The score function $S(\theta)$ is the derivative of the log likelihood with respect to θ,

$$S(\theta) = \frac{\partial}{\partial \theta} l(\theta),$$

and the maximum likelihood estimate is the solution to the score equation $S(\theta) = 0$.

The Fisher information $I(\theta)$, which is defined as

$$I(\theta) = -\frac{\partial^2}{\partial\theta^2} l(\theta),$$

give rise to the observed Fisher information ($I(\hat{\theta})$) and the expected Fisher information ($\mathcal{I}(\theta)$).

1.3.1 Statistical Assumptions for Likelihood Ratio Tests

We generally use LRTs to evaluate the significance of terms in the random effects structure, i.e. different nested models are fitted in which the random effects structure is changed.

When testing hypotheses around covariance parameters in an LME model, REML estimation for both models is recommended by (West et al., 2007), as it REML estimation can be shown to reduce the bias inherent in ML estimates of covariance parameters. Conversely, Pinheiro and Bates (1994) advises that testing hypotheses on fixed-effect parameters should be based on ML estimation, and that using REML would not be appropriate in this context.

A general method for comparing nested models fit by maximum likelihood is the ***likelihood ratio test***. This test can be used for models fit by REML (restricted maximum liklihood), but only if the fixed terms in the two models are invariant, and both models have been fit by REML. Otherwise, the argument: method="ML" must be employed (ML = maximum likelihood).

Generally, likelihood ratio tests should be used to evaluate the significance of terms on the random effects portion of two nested models, and should not be used to determine the significance of the fixed effects. A simple way to more reliably test for the significance of fixed effects in an LME model is to use conditional F-tests, which will give the most reliable test of the fixed effects included inthe model.

1.3.2 Nesting: Model Selection Using Likelihood Ratio Tests

The relationship between the respective models presented by Roy (2009) is known as "nesting". Hypotheses can be formulated in the context of a pair of models that have a nesting relationship West et al. (2007). An important step in the process of model selection is to determine, for a given pair of models, if there is a "nesting relationship" between the two. *Model A* to be nested in the reference model, *Model B*, if *Model A* is a special case of *Model B*, or with some specific constraint applied.

One model is said to be *nested* within another model, i.e. the reference model, if it represents a special case of the reference model (Pinheiro and Bates, 1994).

LRTs can be used to test hypotheses about covariance parameters or fixed effects parameters in the context of LMEs.

1.3.3 Relevance of Estimation Methods

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.

Nested LME models, fitted by ML estimation, can be compared using the likelihood ratio test (Lehmann and Romano, 2006). Models fitted using REML estimation can also be compared, but only if both were fitted using REML, and both have the same fixed effects specifications.

Likelihood ratio tests are generally used to test the significance of terms in the random effects structure.

For both REML and ML estimates, the nominal $p-$values for the LRT statistics under a χ^2 distribution with 2 degrees of freedom are much greater than empirical values. A number of ways of dealing with this issues are discussed (**?**, pg.86).

One should be aware that these p-values may be conservative. That is, the reported p-value may be greater than the true p-value for the test and, in some cases, it may

be much greater.(**?**, pg.87).

Pinheiro & Bates (2000; p. 88) argue that Likelihood Ratio Test comparisons of models varying in fixed effects tend to be anticonservative i.e. will see you observe significant differences in model fit more often than you should.

1.3.4 Akaike Information Critierion

Akaike (1974) introduces the Akaike information criterion (AIC), a model selection tool based on the likelihood function. The AIC is a model selection method, assessing how the goodness of fit of a model. It is computed as follows:

$$AIC = -2l_{max} + 2k$$

with l_{max} as the log-likelihood maximum and k as the number of parameters. Given a data set, candidate models are ranked according to their AIC values, with the model having the lowest AIC being considered the best fit.

Additionally nested models may be compared by using the Akaike Information Criterion,(AIC) and the Bayesian Information Criterion (BIC).

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

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.

Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika 52*(3), 345–370.

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).

Chinchilli, V., J. Martel, S. Kumanyika, and T. Lloyd (1996). A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics 52*(1), 341–353.

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.

Demidenko, E. (2004). *Mixed Models: Theory And Application*. Dartmouth College: Wiley Interscience.

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.

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.

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

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of American Statistical Association 72*(358), 320–338.

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. (1950). Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics 21*, 309–310.

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.

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 (2006a). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Chapman and Hall CRC.

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

Lehmann, E. L. and J. P. Romano (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

McCullough, C. and S. Searle (2001). *Generalized , Linear and Mixed Models*. Wiley Interscience.

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

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

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. (2006). Estimating correlation coefficient between two variables with repeated observations using mixed effects models. *Biometric Journal 2*, 286–301.

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. 18929.

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.

Wolfinger, R. and M. O'connell (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of statistical Computation and Simulation 48*(3-4), 233–243.