Contents

Chapter 1

Introduction

In this section, we introduce the LME model, discusss how it can be applied to MCS problems, and how it is desirable in the case of replicate measurements, giving some examples from previous work (i.e. Carstensen et Al, Lai & Shaio, and Roy). Further to that, there will be a demonstration on fitting various types LME models using freely available software.

While the MCS 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.

1.1 LME models in method comparison studies

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 (2009b) 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 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.

Roy (2009b) 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.

The LME model approach has seen increased use as a framework for method comparison studies in recent years (Lai & Shaio, Carstensen and Choudhary as examples)

Linear mixed effects (LME) models can facilitate greater understanding of the potential causes of bias and differences in precision between two sets of measurement.

LAI-SHIAO

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.Lai and Shiao (2005) view the LME Models approach as an natural expansion to the Bland ? Altman method for comparing two measurement methods. Lai and Shiao (2005) is interesting in that it extends the usual method comparison study question. It correctly identifies LME models as a methodoloy that can used to make such questions tractable. Their focus is to explain lack of agreement by means of additional covariates outside the scope of the traditional method comparison problem.

Lai and Shiao (2005) extends the usual method comparison study question. It correctly identifies LME models as a methodoloy that can used to make such questions tractable. 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.

Carstensen

Carstensen et al. (2008) remarks that modern statistical computation, such as that used for LME models, greatly improve the efficiency of calculation compared to previous 'by-hand' methods.

Due to the prevalence of modern statistical software, Carstensen et al. (2008) advocates the adoption of computer based approaches, such as LME models, to method comparison studies. Carstensen et al. (2008) remarks upon 'by-hand' approaches advocated in Bland and Altman (1999) discouragingly, 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.

Carstensen et al. (2008) remarks upon 'by-hand' approaches advocated in Bland and Altman (1999) discouragingly, describing them as tedious, unnecessary and 'outdated'. 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. Rather than using the 'by hand' methods, estimates for required parameters can be gotten directly from output code. Furthermore, using computer approaches removes constraints, such as the need for the design to be perfectly balanced. 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 **?**, **?**, Cook (1986) West

et al. (2007), amongst others. In this chapter various LME approaches to method comparison studies shall be examined.

Additionally LME based approaches may utilise the diagnostic and influence analysis techniques that have been developed in recent times.

Roys uses and LME model approach to provide a set of formal tests for method comparison studies.

1.2 Introduction to LME Models, Fitting LME Models to MCS 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.

Due to computation complexity, linear mixed effects models have not seen widespread use until many well known statistical software applications began facilitating them.

This approach has seen increased use in 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 LME based approaches may utilise the diagnostic and influence analysis techniques that have been developed in recent times.

1.3 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.

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.

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.

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

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

where y is a vector of N observable random variables, β is a vector of p fixed effects, X and Z are $N \times p$ and $N \times q$ known matrices, and b and ϵ are vectors of q and N, respectively, random effects such that $\mathrm{E}(b) = 0,\ \mathrm{E}(\epsilon) = 0$ and

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

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.

1.4 The Linear Mixed Effects Model

A linear mixed effects model is a linear mdoel that combined fixed and random effect terms formulated by Laird and Ware (1982) as follows;

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i$$

- Y_i is the $n \times 1$ response vector
- X_i is the $n \times p$ Model matrix for fixed effects
- β is the $p \times 1$ vector of fixed effects coefficients
- Z_i is the $n \times q$ Model matrix for random effects
- b_i is the $q \times 1$ vector of random effects coefficients, sometimes denoted as u_i
- ϵ is the $n \times 1$ vector of observation errors

The linear mixed effects model is given by

$$Y = X\beta + Zu + \epsilon \tag{1.2}$$

$\mathbf{Y}$ is the vector of n observations, with dimension $n \times 1$. $\mathbf{b}$ is a vector of fixed p effects, and has dimension $p \times 1$. It is composed of coefficients, with the first element being the population mean. $\mathbf{X}$ is known as the design 'matrix', model matrix for fixed effects, and comprises 0s or 1s, depending on whether the relevant fixed effects have any effect on the observation is question. $\mathbf{X}$ has dimension $n \times p$. $\mathbf{e}$ is the vector of residuals with dimension $n \times 1$.

The random effects models can be specified similarly. $\mathbf{Z}$ is known as the 'model matrix for random effects', and also comprises 0s or 1s. It has dimension $n \times q$. $\mathbf{u}$ is a vector of random q effects, and has dimension $q \times 1$.

1.4.1 Statement of the LME model

These models are used when there are both fixed and random effects that need to be incorporated into a model.

Fixed effects usually correspond to experimental treatments for which one has data for the entire population of samples corresponding to that treatment.

Random effects,on the other hand, are assigned in the case where we have measurements on a group of samples, and those samples are taken from some larger sample pool, and are presumed to be representative.

As such, linear mixed effects models treat the error for fixed effects differently than the error for random effects.

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

$$\mathbf{V} = \mathrm{Var}(\mathbf{Xb} + \mathbf{Zu} + \mathbf{e}) \tag{1.3}$$

$$\mathbf{V} = \mathrm{Var}(\mathbf{Xb}) + \mathrm{Var}(\mathbf{Zu}) + var(\mathbf{e})) \tag{1.4}$$

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

By letting $\mathrm{var}(u) = G$ (i.e $\mathbf{u}$ $N(0, \mathbf{G})$), this becomes ZGZ^T. This specifies the covariance due to random effects. The residual covariance matrix $var(e)$ is denoted as R, ($\mathbf{e} \sim N(0, \mathbf{R})$). 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;

$$\mathbf{V} = ZGZ^T + \mathbf{R} \tag{1.5}$$

The best linear unbiased predictor (BLUP) is used to estimating random effects, i.e to derive $\mathbf{u}$. The best linear unbiased estimator (BLUE) is used to estimate the fixed effects, $\mathbf{b}$. They were formulated in a paper by Henderson et al. (1959), which provides the derivations of both. Inferences about fixed effects have come to be called 'estimates', whereas inferences about random effects have come be called 'predictions'. hence the naming of BLUP is to reinforce distinction between the two , but it is essentially the same principal involved in both cases (GK, 1991). The BLUE of $\mathbf{b}$, and the BLUP of $\mathbf{u}$ can be shown to be;

$$\hat{b} = (X^T V^{-1} X)^{-1} X^T V^{-1} y \tag{1.6}$$

$$\hat{u} = GZ^T V^{-1} (y - X\hat{b}) \tag{1.7}$$

The practical application of both expressions requires that the variance components be known. An estimate for the variance components must be derived to either maximum likelihood (ML) or more commonly restricted maximum likelihood (REML).

Importantly calculations based on the above formulae require the calculation of the inverse of $\mathbf{V}$. In simple examples V^{-1} is a straightforward calculation, but with higher dimensions it becomes a very complex calculation.

1.5 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.

The likelihood function ($L(\theta)$)is a fundamental concept in statistical inference. It indicates how likely a particular population is t o 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 ($\ell(\theta)$).

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.

Likelihood estimation techniques

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

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.

1.6 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 D 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

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

1.7 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

$$\left| \begin{array}{cc} D & 0 \\ 0 & \Sigma \end{array} \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix}' \begin{pmatrix} D & 0 \\ 0 & \Sigma \end{pmatrix}^{-1} \begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix} \right\}, \quad (1.8)$$

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

$$(y - X\beta - Zb)'\Sigma^{-1}(y - X\beta - Zb) + b'D^{-1}b. \quad (1.9)$$

This leads to the mixed model equations

$$\begin{pmatrix} X'\Sigma^{-1}X & X'\Sigma^{-1}Z \\ Z'\Sigma^{-1}X & X'\Sigma^{-1}X + D^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X'\Sigma^{-1}y \\ Z'\Sigma^{-1}y \end{pmatrix}. \quad (1.10)$$

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. **?** 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 (1.10) 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 (1.9) 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.

1.8 Repeated measurements in LME models

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.

1.8.1 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 time point, then the 2 responses at the second time point, 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 $\boldsymbol{\Omega}$.

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.

$$\boldsymbol{y}_i = (y_i^{A1}, y_i^{B1}, y_i^{A2}, y_i^{B2}, y_i^{A3}, y_i^{B3})\prime \tag{1.11}$$

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

$$\boldsymbol{y_i} = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i} + \boldsymbol{\epsilon_i} \tag{1.12}$$

$$\boldsymbol{b_i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}) \tag{1.13}$$

$$\boldsymbol{\epsilon_i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R_i}) \tag{1.14}$$

$\boldsymbol{\beta}$ is a three dimensional vector containing the fixed effects. $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)\prime$. β_2 is usually set to zero. Consequently $\boldsymbol{\beta}$ is the solutions of the means of the two methods, i.e. $E(\boldsymbol{y_i} = \boldsymbol{X_i\beta}$. The variance covariance matrix $\boldsymbol{D}$ is a general 2×2 matrix, while $\boldsymbol{R_i}$ is a $2n_i \times 2n_i$ matrix.

1.9 Decomposition of the response covariance matrix

The variance covariance structure can be re-expressed in the following form,

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

$\boldsymbol{\Omega_i}$ can be expressed as

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

The notation $\dim_{n_i}$ means an $n_i \times n_i$ diagonal block.

$\boldsymbol{R_i}$ 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 **?** use the identity matrix, with dimensions $n_i \times n_i$ as the formulation for $\boldsymbol{V}$. Roy (2009a) 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 (2009a) indicate its use.

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 $\boldsymbol{\Lambda}$.

1.9.1 Correlation terms

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

$$\boldsymbol{D} = \begin{pmatrix} \sigma_A^2 \rho_A & \sigma_A \sigma_b \rho_{AB} \delta \\ \sigma_A \sigma_b \rho_{AB} \delta & \sigma_B^2 \rho_B \end{pmatrix} \tag{1.15}$$

$$\boldsymbol{\Lambda} = \begin{pmatrix} \sigma_A^2 (1 - \rho_A) & \sigma_{AB} (1 - \delta) \\ \sigma_{AB} (1 - \delta) & \sigma_B^2 (1 - \rho_B) \end{pmatrix}. \tag{1.16}$$

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

Chapter 2

LME Model Specification

Model Terms (Roy 2009)

It is important to note the following characteristics of this model.

Let the number of replicate measurements on each item i for both methods be n_i, hence $2 \times n_i$ responses. However, it is assumed that there may be a different number of replicates made for different items. Let the maximum number of replicates be p. An item will have up to $2p$ measurements, i.e. $\max(n_i) = 2p$.

Later on $\boldsymbol{X}_i$ will be reduced to a 2×1 matrix, to allow estimation of terms. This is due to a shortage of rank. The fixed effects vector can be modified accordingly.

$\boldsymbol{Z}_i$ is the $2n_i \times 2$ model matrix for the random effects for measurement methods on item i.

$\boldsymbol{b}_i$ is the 2×1 vector of random-effect coefficients on item i, one for each method.

$\boldsymbol{\epsilon}$ is the $2n_i \times 1$ vector of residuals for measurements on item i.

$\boldsymbol{G}$ is the 2×2 covariance matrix for the random effects.

$\boldsymbol{R}_i$ is the $2n_i \times 2n_i$ covariance matrix for the residuals on item i.

The expected value is given as $\mathrm{E}(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$. (Hamlett et al., 2004)

The variance of the response vector is given by $\text{Var}(\boldsymbol{y}_i) = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R}_i$ (Hamlett et al., 2004).

The maximum likelihood estimate of the between-subject variance covariance matrix of two methods is given as D. The estimate for the within-subject variance covariance matrix is $\hat{\Sigma}$. The estimated overall variance covariance matrix 'Block Ω_i' is the addition of $\hat{D}$ and $\hat{\Sigma}$.

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \tag{2.1}$$

$\boldsymbol{b}_i$ is a $m-$dimensional vector comprised of the random effects.

$$\boldsymbol{b}_i = \begin{pmatrix} b_{1i} \\ b_{21} \end{pmatrix} \tag{2.2}$$

$\boldsymbol{V}$ represents the correlation matrix of the replicated measurements on a given method. $\boldsymbol{\Sigma}$ is the within-subject VC matrix.

$\boldsymbol{V}$ and $\boldsymbol{\Sigma}$ are positive definite matrices. The dimensions of $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$ are $3 \times 3(= p \times p)$ and $2 \times 2(= k \times k)$.

It is assumed that $\boldsymbol{V}$ is the same for both methods and $\boldsymbol{\Sigma}$ is the same for all replications.

$\boldsymbol{V} \bigotimes \boldsymbol{\Sigma}$ creates a $6 \times 6(= kp \times kp)$ matrix. $\boldsymbol{R}_i$ is a sub-matrix of this.

2.1 Model Formula

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

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \tag{2.3}$$

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 $\mathrm{E}(b_{mi}) = 0$ with $\mathrm{Var}(b_{mi}) = g_m^2$ and $\mathrm{Cov}(b_{mi}, b_{m'i}) = g_{12}$. The random error term for each response is denoted ϵ_{mir} having $\mathrm{E}(\epsilon_{mir}) = 0$, $\mathrm{Var}(\epsilon_{mir}) = \sigma_m^2$, $\mathrm{Cov}(b_{mir}, b_{m'ir}) = \sigma_{12}$, $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$ and $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{m'ir'}) = 0$. When two methods of measurement are in agreement, there is no significant differences between β_1 and β_2, g_1^2 andg_2^2, and σ_1^2 andσ_2^2.
Here β_0 and β_m are fixed-effect terms representing, respectively, a model intercept and

an overall effect for method m. The model can be reparameterized by gathering the β terms together into (fixed effect) intercept terms $\alpha_m = \beta_0 + \beta_m$. The b_{1i} and b_{2i} terms are correlated random effect parameters having $\mathrm{E}(b_{mi}) = 0$ with $\mathrm{Var}(b_{mi}) = g_m^2$ and $\mathrm{Cov}(b_{1i}, b_{2i}) = g_{12}$.

The random error term for each response is denoted ϵ_{mir} having $\mathrm{E}(\epsilon_{mir}) = 0$, $\mathrm{Var}(\epsilon_{mir}) = \sigma_m^2$, $\mathrm{Cov}(\epsilon_{1ir}, \epsilon_{2ir}) = \sigma_{12}$, $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$ and $\mathrm{Cov}(\epsilon_{1ir}, \epsilon_{2ir'}) = 0$. Two methods of measurement are in complete agreement if the null hypotheses $\mathrm{H}_1 \colon \alpha_1 = \alpha_2$ and $\mathrm{H}_2 \colon \sigma_1^2 = \sigma_2^2$ and $\mathrm{H}_3 \colon g_1^2 = g_2^2$ hold simultaneously. Roy (2009a) uses a Bonferroni correction to control the familywise error rate for tests of $\{\mathrm{H}_1, \mathrm{H}_2, \mathrm{H}_3\}$ and account for difficulties arising due to multiple testing.

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.

Chapter 3

Introduction to Roy's Procedure

3.1 Roy's Approach

Roy (2009b) 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. 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, 2009b).

Roy (2009b) 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. She provides three tests of hypothesis appropriate for evaluating the agreement between the two methods of measurement under this sampling scheme.

For the purposes of comparing two methods of measurement, Roy (2009b) presents a framework that utilizes 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. Roy (2009b) proposes a suite of hypothesis tests for assess-

ing the agreement of two methods of measurement, when replicate measurements are obtained for each item, using a LME approach. (An item would commonly be a patient). Two methods of measurement are in complete agreement if the null hypotheses $H_1 : \alpha_1 = \alpha_2$ and $H_2 : \sigma_1^2 = \sigma_2^2$ and $H_3 : g_1^2 = g_2^2$ hold simultaneously. Roy (2009b) 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.

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 **?**, it is assumed that the replicates are linked over time. However it is easy to modify to the unlinked case.

Roy (2009b) 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. The well-known "Limits of Agreement", as developed by Bland and Altman (1986) are easily computable using the LME framework, proposed by Roy. While we will not be considering this analysis, a demonstration will be provided in the example.

These tests consider null hypotheses that assume: absence of inter-method bias; equality of between-subject variabilities of the two methods; equality of within-subject variabilities of the two methods. By inter-method bias we mean that a systematic difference exists between observations recorded by the two methods. 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.

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.

Importantly Roy (2009b) 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,

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

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 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.

Roy also integrates H_2 and H_3 into a single testable hypothesis H_4: $\omega_1^2 = \omega_2^2$, where $\omega_m^2 = \sigma_m^2 + g_m^2$ represent the overall variability of method m. 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.

3.2 Replicate measurements in Roy's paper

Roy (2009b) 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.

3.3 Model Set Up

Roy (2009b) 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. 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.

Overall variability between the two methods (Ω) is sum of between-subject (D) and within-subject variability (Σ),

$$\text{Block } \mathbf{\Omega}_i = \begin{bmatrix} d_1^2 & d_{12} \\ d_{12} & d_2^2 \end{bmatrix} + \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

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 **?**, it is assumed that the replicates are linked over time. However it is easy to modify to the unlinked case.

Chapter 4

Model Specification

4.1 Model Specification for Roy's Hypotheses Tests

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.

It is assumed that $\boldsymbol{b}_i \sim N(0, \boldsymbol{G})$, $\boldsymbol{\epsilon}_i$ 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.

The random effects are assumed to be distributed as $\boldsymbol{b}_i \sim \mathcal{N}_2(0, \boldsymbol{G})$. $\boldsymbol{G}$ is the variance covariance matrix for the random effects $\boldsymbol{b}$. i.e. between-item sources of variation. The between-item variance covariance matrix $\boldsymbol{G}$ is constructed as follows:

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

The distribution of the random effects is described as $\boldsymbol{b}_i \sim N(0, \boldsymbol{G})$. 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.

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

It is important to note that no special assumptions about the structure of $\boldsymbol{G}$ are made. An example of such an assumption would be that $\boldsymbol{G}$ is the product of a scalar value and the identity matrix. It is assumed that $\boldsymbol{b}_i \sim N(0, \boldsymbol{G})$, $\boldsymbol{\epsilon}_i$ 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 $\boldsymbol{G}$ and $\boldsymbol{R}_i$ will be discussed in due course.

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

$$\boldsymbol{G} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix}$$

It is important to note that no special assumptions about the structure of $\boldsymbol{G}$ are made. An example of such an assumption would be that $\boldsymbol{G}$ 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 $\boldsymbol{G}$ and $\boldsymbol{R_i}$.

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 $\boldsymbol{G}$ 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 both 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.

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

$\boldsymbol{R}_i$ 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 $\boldsymbol{G}$ 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 ,

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

$\boldsymbol{R}_i$ is the variance covariance matrix for the residuals, i.e. the within-item sources of

variation between both methods. The matrix of random errors $\boldsymbol{\epsilon}_i$ is distributed as $\mathcal{N}_2(0, \boldsymbol{R}_i)$.

4.2 G Component

$\boldsymbol{G}$ is the variance covariance matrix for the random effects $\boldsymbol{b}$. i.e. between-item sources of variation.

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

The distribution of the random effects is described as $\boldsymbol{b}_i \sim N(0, \boldsymbol{G})$. 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.

4.3 R Component

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 $\boldsymbol{G}$ and $\boldsymbol{R_i}$.

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

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

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

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. This may be due to due to the disagreement in either between-item variabilities or within-item variabilities, or both. Roy (2009b) allows for a formal test of each.

The distribution of the random effects is described as $\boldsymbol{b}_i \sim N(0, \boldsymbol{G})$. 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. Both covariance matrices can be written as follows;

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

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

4.4 Hamlett

Hamlett et al. (2004) shows that $\boldsymbol{R}_i$ can be expressed as $\boldsymbol{R}_i = \boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}$. 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. It is assumed that the within-item variance?covariance matrix $\boldsymbol{\Sigma}$ is the same for all replications. Again it is important to note that no special assumptions are made about the structure of the matrix.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \tag{4.1}$$

Hamlett et al. (2004) shows that $\boldsymbol{R}_i$ can be expressed as $\boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}$. The covariance matrix has the same structure for all items, except for dimension, which depends on the number of replicates. The 2×2 block diagonal Block-$\boldsymbol{\Omega}_i$ represents the covariance matrix between two methods, and is the sum of $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$.

$$\text{Block-}\boldsymbol{\Omega}_i = \begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

Hamlett et al. (2004) shows that $\boldsymbol{R}_i$ can be expressed as $\boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}$. The covariance matrix has the same structure for all items, except for dimension, which depends on the number of replicates. The 2×2 block diagonal Block-$\boldsymbol{\Omega}_i$ represents the covariance matrix between two methods, and is the sum of $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$.

$$\text{Block-}\boldsymbol{\Omega}_i = \begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

The variance of case-wise difference in measurements can be determined from Block-$\boldsymbol{\Omega}_i$. Hence limits of agreement can be computed.

4.5 For Expository Purposes

For expository purposes consider the case where each item provides three replicates by each method. Then in matrix notation the model has the structure

$$\boldsymbol{y}_i = \begin{pmatrix} y_{1i1} \\ y_{2i1} \\ y_{1i2} \\ y_{2i2} \\ y_{1i3} \\ y_{2i3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} + \begin{pmatrix} \epsilon_{1i1} \\ \epsilon_{2i1} \\ \epsilon_{1i2} \\ \epsilon_{2i2} \\ \epsilon_{1i3} \\ \epsilon_{2i3} \end{pmatrix}.$$

The between item variance covariance $\boldsymbol{G}$ is as before, while the within item variance covariance is given as

$$\boldsymbol{G} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix}$$

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

Assumptions made on the structures of $\boldsymbol{G}$ and $\boldsymbol{R}_i$ will be discussed in due course.

4.6 Overall Variability

The overall variability between the two methods is the sum of between-item variability $\boldsymbol{G}$ and within-item variability $\boldsymbol{\Sigma}$. Roy (2009b) denotes the overall variability as Block - $\boldsymbol{\Omega}_i$. The overall variation for methods 1 and 2 are given by

$$\begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

The variance of case-wise difference in measurements can be determined from Block-$\mathbf{\Omega}_i$. 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 - $\mathbf{\Omega}_i$ matrix. 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 (2009b) allows for a formal test of each.

4.7 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.

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

Chapter 5

Roy Testing

5.1 Agreement Criteria

Roy sets out three conditions 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. Should both the second and third conditions be fulfilled, then the overall variabilities of both methods would be equal. Roy additionally uses the overall correlation coefficient to provide extra information about the comparison, with a minimum of 0.82 being required.

Roy's method considers two methods to be in agreement if three conditions are met.

- 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.

Roy (2009b) 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. 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

Two methods of measurement can be said to be in agreement if there is no significant difference between in three key respects.

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 (2009b) allows for a formal test of each.

Three tests of hypothesis are provided, appropriate for evaluating the agreement between the two methods of measurement under this sampling scheme. These tests consider null hypotheses that assume: absence of inter-method bias; equality of between-subject variabilities of the two methods; equality of within-subject variabilities of the

two methods. By inter-method bias we mean that a systematic difference exists between observations recorded by the two methods.

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.

5.2 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.

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.

Bias is determinable by examination of the 't-table'. Estimate for both methods are given, and the bias is simply the difference between the two. Because the `R` implementation does not account for an intercept term, a $p-$value is not given. Should a $p-$value be required specifically for the bias, and simple restructuring of the model is

required wherein an intercept term is included. Output from a second implementation will yield a $p-$value.

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 (2009a) 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

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

5.3 Variability Tests

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

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

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

5.4 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 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 $\boldsymbol{G}$ 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.

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

Unstructured

Assumes that each variance and covariance is unique. Each trial has its own variance (e.g. s12 is the variance of trial 1) and each pair of trials has its own covariance (e.g. s21 is the covariance of trial 1 and trial2). This structure is illustrated by the half matrix below.

Autoregressive

Another common covariance structure which is frequently observed in repeated measures data is an autoregressive structure, which recognizes that observations which are more proximate are more correlated than measures that are more distant.

5.5 Roy's Candidate Models : Testing Procedures

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

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

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.

Four candidates models are fitted to the data. These models are similar to one another, but for the imposition of equality constraints.

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

The tests are implemented by fitting a four variants of a specific LME model to the

data. For the purpose of comparing models, one of the models acts as a reference model while the three other variant are nested models that introduce equality constraints to serves as null hypothesis cases. The methodology uses a linear mixed effects regression fit using a combination of symmetric and compound symmetry (CS) correlation structure the variance covariance matrices.

Other important aspects of the method comparison study are consequent. The limits of agreement are computed using the results of the reference model.

5.6 Hypothesis Testing

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

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.

5.7 Roy's hypothesis tests : Roy's variability tests

For the purposes of method comparison, Roy presents a methodology utilising linear mixed effects model. The formulation contains a Kronecker product covariance structure in a doubly multivariate setup. This methodology provides for the formal testing of inter-method bias, between-subject variability and within-subject variability of two methods. 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.

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, Roy proposes a series of three tests on the variance components of an LME model. For these tests, four candidate models are constructed. The difference 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. The first model is compared against each of three other models successively.

$$\mathrm{H}_2 : g_1^2 = g_2^2$$

$$\mathrm{K}_2 : g_1^2 \neq g_2^2$$

and a formal test for the equality of within-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.

5.8 Correlation coefficient

These 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 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 (2009a) 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.

5.9 Roy's variability tests

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.

Other important aspects of the method comparison study are consequent. The limits of agreement are computed using the results of the first model.

The methodology uses a linear mixed effects regression fit using compound symmetry (CS) correlation structure on $\mathbf{V}$.

$\Lambda = \frac{\max_{H_0} L}{\max_{H_1} L}$

5.10 Using LME for method comparison

Due to the prevalence of modern statistical software, Carstensen et al. (2008) advocates the adoption of computer based approaches, such as LME models, to method comparison studies. Carstensen et al. (2008) remarks upon 'by-hand' approaches advocated in Bland and Altman (1999) discouragingly, 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 constraints associated with 'by-hand' approaches, such as the need for the design to be perfectly balanced.

5.10.1 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.

$$H_0 : \ d_A = d_B$$
$$H_A : \ d_A \neq d_B$$

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.

5.10.2 Variability test 2

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

$$H_0 : \ \lambda_A = \lambda_B$$
$$H_A : \ \lambda_A = \lambda_B$$

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 (2009a) includes the coefficients of repeatability for both methods.

5.10.3 Variability test 3

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

$$H_0 : \ \sigma_A = \sigma_B$$
$$H_A : \ \sigma_A = \sigma_B$$

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 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 (2009a) includes the coefficients of repeatability for both 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.

5.10.4 Variability test 3 - Omnibus Test

The maximum likelihood estimate of the between-subject variance covariance matrix of two methods is given as D. The estimate for the within-subject variance covariance matrix is $\hat{\Sigma}$. The estimated overall variance covariance matrix 'Block Ω_i' is the addition of $\hat{D}$ and $\hat{\Sigma}$.

$$\begin{pmatrix} \omega_e^2 & \omega^{en} \\ \omega_{en} & \omega_n^2 \end{pmatrix} = \begin{pmatrix} \psi_e^2 & \psi^{en} \\ \psi_{en} & \psi_n^2 \end{pmatrix} + \begin{pmatrix} \sigma_e^2 & \sigma^{en} \\ \sigma_{en} & \sigma_n^2 \end{pmatrix} \tag{5.1}$$
$$\begin{pmatrix} \omega_2^1 & 0 \\ 0 & \omega_2^2 \end{pmatrix} = \begin{pmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

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 - $\mathbf{\Omega}_i$ matrix. 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 (2009b) allows for a formal test of each.

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \tag{5.2}$$

$$\begin{pmatrix} \omega_e^2 & \omega^{en} \\ \omega_{en} & \omega_n^2 \end{pmatrix} = \begin{pmatrix} \psi_e^2 & \psi^{en} \\ \psi_{en} & \psi_n^2 \end{pmatrix} + \begin{pmatrix} \sigma_e^2 & \sigma^{en} \\ \sigma_{en} & \sigma_n^2 \end{pmatrix} \tag{5.3}$$

5.11 Formal testing for covariances

As it is pertinent to the difference between the two described methodologies, the facilitation of a formal test would be useful. Extending the approach proposed by Roy, 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.

5.12 VC structures

$\mathbf{\Psi}$ is the variance-covariance matrix of the random effects , with 2×2 dimensions.

$$\mathbf{\Psi} = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix} \tag{5.4}$$

There is three alternative structures for $\mathbf{\Psi}$, the diagonal form, the identity form

and the general form.

$$\mathbf{\Psi} = \begin{pmatrix} \psi_1^2 & 0 \\ 0 & \psi_2^2 \end{pmatrix} \quad \text{or} \quad \mathbf{\Psi} = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix} \quad \text{or} \quad \mathbf{\Psi} = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix}$$

Chapter 6

Extending Current Methodologies

6.1 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.

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.

6.2 Conclusion

Carstensen et al. (2008) and Roy (2009a) highlight the need for method comparison methodologies suitable for use in the presence of replicate measurements. Roy (2009a) 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.

6.3 Testing Procedures

Roy's methodology requires the construction of four candidate models. 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 probability distribution of the test statistic can be approximated by a chi-square distribution with (ν_1 - ν_2) degrees of freedom, where ν_1 and ν_2 are the degrees of freedom of models 1 and 2 respectively.

Likelihood ratio tests are very simple to implement in R, simply use the 'anova()' commands. Sample output will be given for each variability test. The likelihood ratio test is the procedure used to compare the fit of two models. For each candidate model, the '-2 log likelihood' ($M2LL$) is computed. The test statistic for each of the three hypothesis tests is the difference of the $M2LL$ for each pair of models. 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.

$$-2 \ln \Lambda_d = [M2LL \text{ under } H_0 \text{ model}] - [M2LL \text{ under } H_A \text{ model}] \tag{6.1}$$

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

$$\nu = [\text{ LRT df under } H_0 \text{ model}] - [\text{ LRT df under } H_A \text{ model}] \tag{6.2}$$

Chapter 7

Likelihood Ratio Tests

7.1 Likelihood

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.

The likelihood function 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.

- Maximum likelihood (ML) estimation is a method of obtaining parameter estimates by optimizing the likelihood function. The likelihood function is constructed as a function of the parameters in the specified model.
- Restricted maximum likelihood (REML) is an alternative methods of computing parameter estimated. REML is often preferred to ML because it produces unbiased estimates of covariance parameters by taking into account the loss of degrees of freedom that results from estimating the fixed effects in $\boldsymbol{\beta}$.

A general method for comparing nested models fitted by ML is the ***likelihood ratio test*** (Cite: Lehmann 1986). 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. Each of these three test shall be examined in more detail shortly.

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.

A general method for comparing models with a nesting relationship is the likelihood ratio test (LRTs). 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).

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.

7.2 Likelihood Ratio Tests in Roy's Analysis

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.

7.3 Nesting: Model Selection Using Likelihood Ratio Tests

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.

We define Model A to be "nested" in Model B if Model A is a special case of Model B, i.e. Model B with a 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).

Hypotheses can be formulated in the context of a pair of models that have a nesting relationship West et al. (2007).

LRTs are a class of tests used to compare the value of likelihood functions for two models defining a hypothesis to be tested (i.e. the nested and reference model).

The relationship between the respective models presented by Roy (2009a) is known as "nesting". A 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.

7.4 Statistical Assumptions for Likelihood Ratio Tests

If k_i is the number of parameters to be estimated in model i, then the asymptotic, or "large sample", distribution of the LRT statistic, under the null hypothesis that the restricted model is adequate, is a χ^2 distribution with $k_2 - k_1$ degrees of freedom (**?**, pg.83).

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.

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.

When testing hypotheses around covariance parameters in an LME model, REML estimation for both models is recommended by West et al. REML estimation can be shown to reduce the bias inherent in ML estimates of covariance parameters (West et al., 2007). Conversely, **?** 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.

7.5 Other material

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).

- Example of a likelihood ratio test used to compare two models:
  `>anova(modelA, modelB)`
- The output will contain a p-value, and this should be used in conjunction with the AIC scores to judge which model is preferred. Lower AIC scores are better.
- 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, as implemented with the simple "anova" function. Example:
  `>anova(modelA)`
  will give the most reliable test of the fixed effects included in model1.

7.5.1 Likelihood Ratio Tests

The relationship between the respective models presented by Roy (2009b) is known as "nesting". A 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.

When testing hypotheses around covariance parameters in an LME model, REML estimation for both models is recommended by West et al. REML estimation can be

shown to reduce the bias inherent in ML estimates of covariance parameters. Conversely, Roy (2009b) 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.

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.

Testing Procedures

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

Likelihood ratio tests are very simple to implement in `R`, simply use the '`anova()`' commands. Sample output will be given for each variability test. The likelihood ratio test is the procedure used to compare the fit of two models. For each candidate model, the '-2 log likelihood' (M2LL) is computed. The test statistic for each of the three hypothesis tests is the difference of the M2LL for each pair of models. 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.

$$-??2ln\Lambda_d = [\text{M2LL under H0 model}] - [\text{M2LL under HA model}] \tag{7.1}$$

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

$$\nu_{=}[\text{LRT df under H0 model}] - [\text{ LRT df under HA model}] \tag{7.2}$$

Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
MCS1	8	4077.5	4111.3	-2030.7			
MCS2	7	4075.6	4105.3	-2030.8	1 vs 2	0.15291	0.6958

7.6 LRTs for covariance parameters

[cite: West et al] When testing hypotheses around covariance parameters in an LME model, REML estimation for both models is recommended by West et al. REML estimation can be shown to reduce the bias inherent in ML estimates of covariance parameters [cite: Morrel98]

7.7 Test Statistic for Likelihood Ratio Tests

The likelihood ratio test is the procedure used to compare the fit of two models. For each candidate model, the '-2 log likelihood' ($M2LL$) is computed. The test statistic for each of the three hypothesis tests is the difference of the $M2LL$ for each pair of models.

The test statistic for the likelihood ratio test is the difference of the log-likelihood functions, multiplied by -2. The test statistic for the LRT is the difference of the log-likelihood functions, multiplied by -2. L= - 2ln is approximately distributed as 2 under H_0 for large sample size and under the normality assumption.

$$-2 \ln \Lambda_d = [M2LL \text{ under } H_0 \text{ model}] - [M2LL \text{ under } H_A \text{ model}] \tag{7.3}$$

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

$$\nu = [\text{ LRT df under } H_0 \text{ model}] - [\text{ LRT df under } H_A \text{ model}] \tag{7.4}$$

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 significance of the likelihood ratio test can be found by comparing it to the χ^2 distribution, with the appropriate degrees of freedom.

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 ($\mathcal{I}(\theta)$).

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.

7.8 Relevance of Estimation Methods

When testing hypotheses around covariance parameters in an LME model, REML estimation for both models is recommended by West et al. REML estimation can be shown to reduce the bias inherent in ML estimates of covariance parameters (West et al., 2007). Conversely, **?** 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.

Nested LME models, fitted by ML estimation, can be compared using the likelihood ratio test [Lehmann (1986)]. 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.

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 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.

A general method for comparing nested models fitted by ML is the ***likelihood ratio test*** (Cite: Lehmann 1986). 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.

If k_i is the number of parameters to be estimated in model i, then the asymptotic, or "large sample", distribution of the LRT statistic, under the null hypothesis that the restricted model is adequate, is a χ^2 distribution with $k_2 - k_1$ degrees of freedom (**?**, pg.83).

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.

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.

7.9 Information Criteria

Akaike (1974) introduces the Akaike information criterion (AIC), a model selection tool based on the likelihood function. 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).

When comparing the respective scores for nested models, the model with the smaller score is considered to be the preferable model. ML / REML [Morrell 1998] 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.

Pinheiro and Bates (1994) addresses the issue of treating items as fixed effects. Such a specification is useful only for the specific sample of items, rather than the population of items, where the interest would naturally lie.

Pinheiro and Bates (1994) advises the specification of random effects to correspond to items; treating the item effects as random deviations from the population mean.

7.10 BXC - Model Terms

- Let y_{mir} be the response of method m on the ith subject at the $r-$th replicate.
- Let $\boldsymbol{y}_{ir}$ be the 2×1 vector of measurements corresponding to the $i-$th subject at the $r-$th replicate.

- Let $\boldsymbol{y}_i$ be the $R_i \times 1$ vector of measurements corresponding to the $i-$th subject, where R_i is number of replicate measurements taken on item i.
- Let $\alpha_m i$ be the fixed effect parameter for method for subject i.
- Formally Roy uses a separate fixed effect parameter to describe the true value μ_i, but later combines it with the other fixed effects when implementing the model.
- Let u_{1i} and u_{2i} be the random effects corresponding to methods for item i.
- $\boldsymbol{\epsilon}_i$ is a n_i-dimensional vector comprised of residual components. For the blood pressure data $n_i = 85$.
- $\boldsymbol{\beta}$ is the solutions of the means of the two methods. In the LME output, the bias ad corresponding t-value and p-values are presented. This is relevant to Roy's first test.

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 (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).

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

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.

GK, R. (1991). That blups are a good thing: The estimation of random effects. *Statistical Science 6*(1), 15–32.

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.

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

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

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

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.