

# Contents

<b>1</b>	<b>Linear Mixed effects Models</b>	<b>4</b>
1.1	(good)Linear Mixed effects Models . . . . .	4
1.1.1	(good)Estimation . . . . .	6
1.2	(good)Repeated measurements in LME models . . . . .	11
1.2.1	(good)Formulation of the response vector . . . . .	11
1.2.2	(good)Decomposition of the response covariance matrix . . . . .	12
1.2.3	(good)Correlation terms . . . . .	13
1.3	(good)Using LME for method comparison . . . . .	15
1.3.1	(good)Roy's methodology . . . . .	15
1.3.2	(good)Correlation . . . . .	16
1.3.3	(good)Variability test 1 . . . . .	17
1.3.4	(good)Variability test 2 . . . . .	17
1.3.5	(good)Variability test 3 . . . . .	18
1.3.6	(good)Demonstration of Roy's testing . . . . .	18
1.4	(good)Implementation in R . . . . .	21
1.5	(good)Extension of Roy's methodology . . . . .	24
<b>2</b>	<b>Linear Mixed effects Models</b>	<b>26</b>
2.1	Linear Mixed effects Models . . . . .	26
2.1.1	Laird Ware Model . . . . .	27
2.1.2	Henderson's equations . . . . .	35

2.1.3	Likelihood and estimation . . . . .	36
2.1.4	Algorithms . . . . .	37
2.1.5	Estimation of Fixed and Random Effects . . . . .	39
2.2	LME models in method comparison studies . . . . .	40
2.2.1	Agreement Criteria . . . . .	41
2.2.2	Fitting LME Models to Method Comparison Data . . . . .	43
2.2.3	Roy's Approach . . . . .	43
2.2.4	Replicate measurements in Roy's paper . . . . .	44
2.2.5	Specifying the Models . . . . .	44
2.2.6	Model Specification for Roy's Hypotheses Tests . . . . .	46
2.2.7	LME Model Specification . . . . .	46
2.2.8	Test for inter-method bias . . . . .	47
2.2.9	Roy's hypothesis tests : Roy's variability tests . . . . .	48
2.2.10	Variance Covariance Matrices . . . . .	48
2.2.11	Variability Tests . . . . .	50
2.2.12	Computing Limits of Agreement . . . . .	53
2.2.13	Formal testing for covariances (Off-Diagonal Components in Roy's Model) . . . . .	53
2.2.14	Correlation coefficient . . . . .	54
2.2.15	Extension of Roy's methodology . . . . .	55
2.2.16	Roy's methodology for single measurements . . . . .	56
2.3	(good)Limits of agreement in LME models . . . . .	56
2.3.1	(good)Linked replicates . . . . .	58
2.4	(good)Conclusion . . . . .	60
	Bibliography . . . . .	60
2.5	Likelihood Ratio Tests . . . . .	61
2.5.1	Statistical Assumptions for Likelihood Ratio Tests . . . . .	62
2.5.2	Nesting: Model Selection Using Likelihood Ratio Tests . . . . .	63
2.5.3	Relevance of Estimation Methods . . . . .	63

2.5.4 Akaike Information Criterion . . . . .	64
--	----

# Chapter 1

## Linear Mixed effects Models

### 1.1 (good)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.

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

The methodology has developed since, including contributions from ?, who extend the use of variance components into linear models, and ?, who introduced the ‘mixed model’ terminology and formally distinguished between mixed and random effects models. ? devised a methodology for deriving estimates for both the fixed effects and the

random effects, using a set of equations that would become known as ‘mixed model equations’ or ‘Henderson’s equations’. LME methodology is further enhanced by Henderson’s later works (1975). The key features of Henderson’s work provide the basis for the estimation techniques.

Wallerstein 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. Wallerstein 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.

Wallerstein 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 (1992). Wallerstein described how to compute LME models in the **S-plus** environment.

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

$$y = X\beta + Zb + \epsilon \quad (1.1)$$

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

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

where  $D$  and  $\Sigma$  are positive definite matrices parameterized by an unknown variance component parameter vector  $\theta$ . 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.

### 1.1.1 (good)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  $\Sigma$ . 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 (2.1), 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

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

### (good)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 ????? offered a more simpler approach of jointly estimating  $\hat{\beta}$  and  $\hat{b}$ . ? made the (ad-hoc) distributional assumptions  $y|b \sim N(X\beta + Zb, \Sigma)$  and  $b \sim N(0, D)$ , and proceeded to maximize the joint density of  $y$  and  $b$

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

with respect to  $\beta$  and  $b$ , which ultimately requires minimizing the criterion

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

This leads to the mixed model equations

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

Using these equations, obtaining the estimates requires the inversion of a matrix of dimension  $p + q \times p + q$ , considerably smaller in size than  $V$ . ? 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  $\Sigma$  are known.

? points out that although ? initially referred to the estimates  $\hat{\beta}$  and  $\hat{b}$  from (2.7) as “joint maximum likelihood estimates”, ? later advised that these estimates should not be referred to as “maximum likelihood” as the function being maximized in (2.6) is a joint density rather than a likelihood function. ? remarks that it is clear that Henderson used joint estimation for computational purposes, without recognizing the theoretical implications.

### (good)Estimation of the fixed parameters

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

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

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

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

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

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

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

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

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

The REML approach does not base estimates on a maximum likelihood fit of all the information, but instead uses a likelihood function derived from a data set, transformed to remove the irrelevant influences (?). 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  $\beta$ . 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.

### **(good) Estimation of the random effects**

The established approach for estimating the random effects is to use the best linear predictor of  $b$  from  $y$ , which for a given  $\beta$  equals  $DZ'V^{-1}(y - X\beta)$ . In practice  $\beta$  is replaced by an estimator such as  $\hat{\beta}$  from (1.6) so that  $\hat{b} = DZ'V^{-1}(y - X\hat{\beta})$ . Pre-multiplying by the appropriate matrices it is straightforward to show that these estimates  $\hat{\beta}$  and  $\hat{b}$  satisfy the equations in (2.7).

### **(good) Algorithms for likelihood function optimization**

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



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

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

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

### (good)The extended likelihood

The desire to have an entirely likelihood-based justification for estimates of random effects, in contrast to Henderson’s equation, has motivated ?, page 429 to define the *extended likelihood*. He remarks “In mixed effects modelling the extended likelihood has been called *h-likelihood* (for hierarchical likelihood) by ?, while in smoothing literature it is known as the *penalized likelihood* (e.g. ? ?).” The extended likelihood can be written  $L(\beta, \theta, b|y) = p(y|b; \beta, \theta)p(b; \theta)$  and adopting the same distributional

assumptions used by ? yields the log-likelihood function

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

Given  $\theta$ , differentiating with respect to  $\beta$  and  $b$  returns Henderson's equations in (2.7).

### (good)The LME model as a general linear model

Henderson's equations in (2.7) can be rewritten  $(T'W^{-1}T)\delta = T'W^{-1}y_a$  using

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

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

## 1.2 (good)Repeated measurements in LME models

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

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

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

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

The matrix  $\boldsymbol{\Sigma}$  represents the variance component matrix between response variables at a given time point  $j$ .

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

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

### 1.2.1 (good)Formulation of the response vector

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

Consider the case where three measurements are taken by both methods  $A$  and  $B$ ,

$\mathbf{y}_i$  is a  $6 \times 1$  random vector describing the  $i$ th subject.

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

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

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i$$

$$\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$$

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_i)$$

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

### 1.2.2 (good)Decomposition of the response covariance matrix

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

$$\text{Cov}(\mathbf{y}_i) = \boldsymbol{\Omega}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i.$$

$\mathbf{R}_i$  can be shown to be the Kronecker product of a correlation matrix  $\mathbf{V}$  and  $\mathbf{\Lambda}$ . The correlation matrix  $\mathbf{V}$  of the repeated measures on a given response variable is assumed to be the same for all response variables. Both ? and ? use the identity matrix, with dimensions  $n_i \times n_i$  as the formulation for  $\mathbf{V}$ . ? 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. ? 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 ? indicate its use.

For the response vector described, ? presents a detailed covariance matrix. A brief summary shall be presented here only. The overall variance matrix is a  $6 \times 6$  matrix composed of two types of  $2 \times 2$  blocks. Each block represents one separate time of measurement.

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

The diagonal blocks are  $\Sigma$ , as described previously. The  $2 \times 2$  block diagonal matrix in  $\mathbf{\Omega}$  gives  $\Sigma$ .  $\Sigma$  is the sum of the between-subject variability  $\mathbf{D}$  and the within subject variability  $\mathbf{\Lambda}$ .

$\mathbf{\Omega}_i$  can be expressed as

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

The notation  $\text{dim}_{n_i}$  means an  $n_i \times n_i$  diagonal block.

### 1.2.3 (good)Correlation terms

? demonstrated how the between-subject and within subject variabilities can be expressed in terms of correlation terms.

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

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

$\rho_A$  describe the correlations of measurements made by the method  $A$  at different times. Similarly  $\rho_B$  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.  $\rho_{AB}$  describes the correlation of measurements taken at the same same time by both methods. The coefficient  $\delta$  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  $\delta$  is simply 1. ? provides a useful graphical depiction of the role of each correlation coefficients.

## 1.3 (good)Using LME for method comparison

Due to the prevalence of modern statistical software, ? advocates the adoption of computer based approaches, such as LME models, to method comparison studies. ? remarks upon ‘by-hand’ approaches advocated in ? 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.

### 1.3.1 (good)Roy’s methodology

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

? 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.

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 ? 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  $D$  and  $\Lambda$  matrices are constructed, using either an unstructured form or a compound symmetry form. To illustrate these differences, consider a generic matrix  $A$ ,

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

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

The first model acts as an alternative hypothesis to be compared against each of three other models, acting as null hypothesis models, successively. The models are compared using the likelihood ratio test. Likelihood ratio tests are a class of tests based on the comparison of the values of the likelihood functions of two candidate models. LRTs can be used to test hypotheses about covariance parameters or fixed effects parameters in the context of LMEs. The test statistic for the likelihood ratio test is the difference of the log-likelihood functions, multiplied by  $-2$ . The probability distribution of the test statistic is approximated by the  $\chi^2$  distribution with  $(\nu_1 - \nu_2)$  degrees of freedom, where  $\nu_1$  and  $\nu_2$  are the degrees of freedom of models 1 and 2 respectively. Each of these three test shall be examined in more detail shortly.

### 1.3.2 (good)Correlation

In addition to the variability tests, Roy advises that it is preferable that a correlation of greater than 0.82 exist for two methods to be considered interchangeable. However if two methods fulfil all the other conditions for agreement, failure to comply with this



one can be overlooked. Indeed Roy demonstrates that placing undue importance to it can lead to incorrect conclusions. ? remarks that current computer implementations only gives overall correlation coefficients, but not their variances. Consequently it is not possible to carry out inferences based on all overall correlation coefficients.

### 1.3.3 (good)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  $\hat{\mathbf{A}}$  has a symmetric form for both models, and will be the same for both.

### 1.3.4 (good)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.

$$H_0 : \lambda_A = \lambda_B$$

$$H_A : \lambda_A \neq \lambda_B$$

This model is performed in the same manner as the first test, only reversing the roles of  $\hat{\mathbf{D}}$  and  $\hat{\mathbf{A}}$ . The null model is constructed a symmetric form for  $\hat{\mathbf{A}}$  while the alternative model uses a compound symmetry form. This time  $\hat{\mathbf{D}}$  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, ? includes the coefficients of repeatability for both methods.

### 1.3.5 (good)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  $\hat{D}$  and  $\hat{A}$  while the alternative model uses a compound symmetry form for both.

### 1.3.6 (good)Demonstration of Roy's testing

Roy provides three case studies, using data sets well known in method comparison studies, to demonstrate how the methodology should be used. The first two examples used are from the ‘blood pressure’ data set introduced by ?. The data set is a tabulation of simultaneous measurements of systolic blood pressure were made by each of two experienced observers (denoted ‘J’ and ‘R’) using a sphygmomanometer and by a semi-automatic blood pressure monitor (denoted ‘S’). Three sets of readings were made in quick succession. Roy compares the ‘J’ and ‘S’ methods in the first of her examples.

The inter-method bias between the two method is found to be 15.62 , with a  $t$ -value of  $-7.64$ , with a  $p$ -value of less than 0.0001. Consequently there is a significant inter-method bias present between methods  $J$  and  $S$ , and the first of the Roy’s three agreement criteria is unfulfilled.

Next, the first variability test is carried out, yielding maximum likelihood estimates of the between-subject variance covariance matrix, for both the null model, in compound symmetry (CS) form, and the alternative model in symmetric (symm) form.

These matrices are determined to be as follows;

$$\hat{\mathbf{D}}_{CS} = \begin{pmatrix} 946.50 & 784.32 \\ 784.32 & 946.50 \end{pmatrix}, \quad \hat{\mathbf{D}}_{Symm} = \begin{pmatrix} 923.98 & 785.24 \\ 785.24 & 971.30 \end{pmatrix}.$$

A likelihood ratio test is perform to compare both candidate models. The log-likelihood of the null model is  $-2030.7$ , and for the alternative model  $-2030.8$ . The test statistic, presented with greater precision than the log-likelihoods, is  $0.1592$ . The  $p$ -value is  $0.6958$ . Consequently we fail to reject the null model, and by extension, conclude that the hypothesis that methods  $J$  and  $S$  have the same between-subject variability. Thus the second of the criteria is fulfilled.

The second variability test determines maximum likelihood estimates of the within-subject variance covariance matrix, for both the null model, in CS form, and the alternative model in symmetric form.

$$\hat{\mathbf{\Lambda}}_{CS} = \begin{pmatrix} 60.27 & 16.06 \\ 16.06 & 60.27 \end{pmatrix}, \quad \hat{\mathbf{\Lambda}}_{Symm} = \begin{pmatrix} 37.40 & 16.06 \\ 16.06 & 83.14 \end{pmatrix}.$$

Again, A likelihood ratio test is perform to compare both candidate models. The log-likelihood of the alternative model model is  $-2045.0$ . As before, the null model has a log-likelihood of  $-2030.7$ . The test statistic is computed as  $28.617$ , again presented with greater precision. The  $p$ -value is less than  $0.0001$ . In this case we reject the null hypothesis of equal within-subject variability. Consequently the third of Roy's criteria is unfulfilled. The coefficient of repeatability for methods  $J$  and  $S$  are found to be  $16.95$  mmHg and  $25.28$  mmHg respectively.

The last of the three variability tests is carried out to compare the overall variabilities of both methods. With the null model the MLE of the within-subject variance covariance matrix is given below. The overall variabilities for the null and alternative models, respectively, are determined to be as follows;

$$\hat{\mathbf{\Sigma}}_{CS} = \begin{pmatrix} 1007.92 & 801.65 \\ 801.65 & 1007.92 \end{pmatrix}, \quad \hat{\mathbf{\Sigma}}_{Symm} = \begin{pmatrix} 961.38 & 801.40 \\ 801.40 & 1054.43 \end{pmatrix},$$

The log-likelihood of the alternative model is  $-2045.2$ , and again, the null model has a log-likelihood of  $-2030.7$ . The test statistic is  $28.884$ , and the  $p$ -value is less than  $0.0001$ . The null hypothesis, that both methods have equal overall variability, is rejected. Further to the second variability test, it is known that this difference is specifically due to the difference of within-subject variabilities.

Lastly, Roy considers the overall correlation coefficient. The diagonal blocks  $\hat{\mathbf{r}}_{\Omega_{ii}}$  of the correlation matrix indicate an overall coefficient of  $0.7959$ . This is less than the threshold of  $0.82$  that Roy recommends.

$$\hat{\mathbf{r}}_{\Omega_{ii}} = \begin{pmatrix} 1 & 0.7959 \\ 0.7959 & 1 \end{pmatrix}$$

The off-diagonal blocks of the overall correlation matrix  $\hat{\mathbf{r}}_{\Omega_{ii'}}$  present the correlation coefficients further to  $?$ .

$$\hat{\mathbf{r}}_{\Omega_{ii'}} = \begin{pmatrix} 0.9611 & 0.7799 \\ 0.7799 & 0.9212 \end{pmatrix}.$$

The overall conclusion of the procedure is that method  $J$  and  $S$  are not in agreement, specifically due to the within-subject variability, and the inter-method bias. The repeatability coefficients are substantially different, with the coefficient for method  $S$  being  $49\%$  larger than for method  $J$ . Additionally the overall correlation coefficient did not exceed the recommended threshold of  $0.82$ .

## 1.4 (good)Implementation in R

To implement an LME model in R, the `nlme` package is used. This package is loaded into the R environment using the library command, (i.e. `library(nlme)`). The `lme` command is used to fit LME models. The first two arguments to the `lme` function specify the fixed effect component of the model, and the data set to which the model is to be fitted. The first candidate model (‘MCS1’) fits an LME model on the data set ‘dat’. The variable ‘method’ is assigned as the fixed effect, with the response variable ‘BP’ (i.e. blood pressure).

The third argument contain the random effects component of the formulation, describing the random effects, and their grouping structure. The `nlme` package provides a set of positive-definite matrices, the `pdMat` class, that can be used to specify a structure for the between-subject variance-covariance matrix for the random effects. For Roy’s methodology, we will use the `pdSymm` and `pdCompSymm` to specify a symmetric structure and a compound symmetry structure respectively. A full discussion of these structures can be found in ?, pg. 158.

Similarly a variety of structures for the within-subject variance-covariance matrix can be implemented using `nlme`. To implement a particular matrix structure, one must specify both a variance function and correlation structure accordingly. Variance functions are used to model the variance structure of the within-subject errors. `varIdent` is a variance function object used to allow different variances according to the levels of a classification factor in the data. A compound symmetry structure is implemented using the `corCompSymm` class, while the symmetric form is specified by `corSymm` class. Finally, the estimation methods is specified as “ML” or “REML”.

The first of Roy's candidate model can be implemented using the following code;

---

```
MCS1 = lme(BP ~ method-1, data = dat,  
random = list(subject=pdSymm(~ method-1)),  
weights=varIdent(form=~1|method),  
correlation = corSymm(form=~1 | subject/obs), method="ML")
```

---

For the blood pressure data used in ?, all four candidate models are implemented by slight variations of this piece of code, specifying either `pdSymm` or `pdCompSymm` in the second line, and either `corSymm` or `corCompSymm` in the fourth line. For example, the second candidate model 'MCS2' is implemented with the same code as MCS1, except for the term `pdCompSymm` in the second line, rather than `pdSymm`.

---

```
MCS2 = lme(BP ~ method-1, data = dat,  
random = list(subject=pdCompSymm(~ method-1)),  
weights = varIdent(form=~1|method),  
correlation = corSymm(form=~1 | subject/obs), method="ML")
```

---

Using this R implementation for other data sets requires that the data set is structured appropriately (i.e. each case of observation records the index, response, method and replicate). Once formatted properly, implementation is simply a case of re-writing the first line of code, and computing the four candidate models accordingly.

To perform a likelihood ratio test for two candidate models, simply use the `anova` command with the names of the candidate models as arguments. The following piece of code implement the first of Roy's variability tests.

---

```
> anova(MCS1,MCS2)
```

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

```
>
```

---

The fixed effects estimates are the same for all four candidate models. The inter-method bias can be easily determined by inspecting a summary of any model. The summary presents estimates for all of the important parameters, but not the complete variance-covariance matrices (although some simple R functions can be written to overcome this). The variance estimates for the random effects for MCS2 is presented below.

---

```
Random effects:
Formula: ~method - 1 | subject
Structure: Compound Symmetry
StdDev Corr
methodJ 30.765
methodS 30.765 0.829
Residual 6.115
```

---

Similarly, for computing the limits of agreement the standard deviation of the differences is not explicitly given. Again, A simple R function can be written to calculate the limits of agreement directly.

## 1.5 (good)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 ?. 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.

# Chapter 2

## Linear Mixed effects Models

? and ? highlight the need for method comparison methodologies suitable for use in the presence of replicate measurements. ? presents a comprehensive methodology for assessing the agreement of two methods, for replicate measurements. This methodology has the added benefit of overcoming the problems of unbalanced data and unequal numbers of replicates. Implementation of the methodology, and interpretation of the results, is relatively easy for practitioners who have only basic statistical training. Furthermore, it can be shown that widely used existing methodologies, such as the limits of agreement, can be incorporated into Roy's methodology.

While the method comparison problem is conventionally poised in the context of two methods of measurements, LME models allow for a straightforward analysis whereby several methods of measurement can be measured simultaneously. However simple models only can only indicate agreement or 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.

### 2.1 Linear Mixed effects Models

A linear mixed effects (LME) model is a statistical model containing both fixed effects and random effects (random effects are also known as variance components). LME

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

? introduced variance components models for use in genetical studies. Whereas an estimate for variance must take a 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 ?, who extend the use of variance components into linear models, and ?, who introduced the ‘mixed model’ terminology and formally distinguished between mixed and random effects models. ? 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 (?????). The key features of Henderson’s work provide the basis for the estimation techniques.

### 2.1.1 Laird Ware Model

? 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 (?). ? described how to compute LME models in the S-plus environment.

Linear mixed effects models (LME) differs from the conventional linear model in that it has both fixed effects and random effects regressors, and coefficients thereof.

The notation provided here is generic, and will be adapted to accord with complex formulations that will be encountered in due course. Using Laird-Ware form, the LME model is commonly described in matrix form,

$$Y = X\beta + Zb + \epsilon \quad (2.1)$$

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

### Formulation of the response vector

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

Consider the case where three measurements are taken by both methods  $A$  and  $B$ ,  $\mathbf{y}_i$  is a  $6 \times 1$  random vector describing the  $i$ th subject.

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

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

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i$$

$$\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$$

$$\epsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_i)$$

Information on the fixed effects are contained in a three dimensional vector  $\beta = (\beta_0, \beta_1, \beta_2)'$ . For computational purposes  $\beta_2$  is conventionally set to zero. Consequently

$\beta$  is the solutions of the means of the two methods, i.e.  $E(\mathbf{y}_i) = \mathbf{X}_i\beta$ . The variance covariance matrix  $\mathbf{D}$  is a general  $2 \times 2$  matrix, while  $\mathbf{R}_i$  is a  $2n_i \times 2n_i$  matrix.

The LME model can be written

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i,$$

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

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

$$\Omega_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i.$$

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

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

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

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

## The Variance Covariance Matrix

The variance matrix of  $\mathbf{Y}$ , denoted  $\mathbf{V}$ , is an  $n \times n$  matrix that can be expressed as follows;

$$\mathbf{V} = \text{Var}(\mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{b} + \mathbf{e}) \quad (2.2)$$

$$\mathbf{V} = \text{Var}(\mathbf{X}\mathbf{b}) + \text{Var}(\mathbf{Z}\mathbf{b}) + \text{Var}(\mathbf{e}) \quad (2.3)$$

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

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

where  $D$  and  $\Sigma$  are positive definite matrices parameterized by an unknown variance component parameter vector  $\theta$ . The variance-covariance matrix for the vector of observations  $y$  is given by  $V = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \Sigma$ . This implies  $y \sim (X\beta, V) = (X\beta, \mathbf{Z}\mathbf{D}\mathbf{Z}' + \Sigma)$ .

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

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

$$\mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}^T + \mathbf{R} \quad (2.4)$$

## Decomposition of the response covariance matrix

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

$$\text{Cov}(y_i) = \Omega_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i.$$

$\mathbf{R}_i$  can be shown to be the Kronecker product of a correlation matrix  $\mathbf{V}$  and  $\mathbf{\Lambda}$ . The correlation matrix  $\mathbf{V}$  of the repeated measures on a given response variable is assumed to be the same for all response variables. It is important to note that no special assumptions about the structure of  $\mathbf{D}$  are made. An example of such an assumption would be that  $\mathbf{D}$  is the product of a scalar value and the identity matrix. Both ? and ? use the identity matrix, with dimensions  $n_i \times n_i$  as the formulation for  $\mathbf{V}$ . ? 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. ? 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 ? indicate its use.

The matrix of random errors  $\boldsymbol{\epsilon}_i$  is distributed as  $\mathcal{N}_2(0, \mathbf{R}_i)$ . ? 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 Kronecker product of an  $n_i \times n_i$  identity matrix and the partial within-item variance covariance matrix  $\boldsymbol{\Sigma}$ , i.e.  $\mathbf{R}_i = \mathbf{I}_{n_i} \otimes \boldsymbol{\Sigma}$ .

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

where  $\sigma_1^2$  and  $\sigma_2^2$  are the within-subject variances of the respective methods, and  $\sigma_{12}$  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  $\mathbf{D}$  and  $\mathbf{R}_i$ .

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

? shows that  $\mathbf{R}_i$  can be expressed as  $\mathbf{R}_i = \mathbf{I}_{n_i} \otimes \mathbf{\Sigma}$ . The partial within-item variance-covariance matrix of two methods at any replicate is denoted  $\mathbf{\Sigma}$ , where  $\sigma_1^2$  and  $\sigma_2^2$  are the within-subject variances of the respective methods, and  $\sigma_{12}$  is the within-item covariance between the two methods.

For the response vector described, ? presents a detailed covariance matrix. A brief summary shall be presented here only. The overall variance matrix is a  $6 \times 6$  matrix composed of two types of  $2 \times 2$  blocks. Each block represents one separate time of measurement.

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

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

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

? 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  $\mathbf{\Sigma}$ , i.e.  $\mathbf{R}_i = \mathbf{I}_{n_i} \otimes \mathbf{\Sigma}$ .



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

where  $\sigma_1^2$  and  $\sigma_2^2$  are the within-subject variances of the respective methods, and  $\sigma_{12}$  is the within-item covariance between the two methods.

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

### Repeated measurements in LME models

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

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

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

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

The matrix  $\mathbf{\Sigma}$  represents the variance component matrix between response variables at a given time point  $j$ .

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

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

## Correlation terms

? demonstrated how the between-subject and within subject variabilities can be expressed in terms of correlation terms.

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

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

$\rho_A$  describe the correlations of measurements made by the method  $A$  at different times. Similarly  $\rho_B$  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.  $\rho_{AB}$  describes the correlation of measurements taken at the same same time by both methods. The coefficient  $\delta$  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  $\delta$  is simply 1. ? provides a useful graphical depiction of the role of each correlation coefficients.

? 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  $\rho_{xy}$  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 (2.8) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items  $N$ , whereas the model in (??) requires  $N + 2$  fixed effects.

Allocating fixed effects to each item  $i$  by (??)(**OFF CHAPTER**) accords with earlier work on comparing methods of measurement, such as ?. 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, ? uses a more intuitive approach, treating the observations as a random sample population, and allocating random effects accordingly.

### 2.1.2 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 this problem ????? offered a more simpler approach of jointly estimating  $\hat{\beta}$  and  $\hat{b}$ . ? made the (ad-hoc) distributional assumptions  $y|b \sim N(X\beta + Zb, \Sigma)$  and  $b \sim N(0, D)$ , and proceeded to maximize the joint density of  $y$  and  $b$

$$\left| \begin{matrix} D & 0 \\ 0 & \Sigma \end{matrix} \right|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix}' \begin{pmatrix} D & 0 \\ 0 & \Sigma \end{pmatrix}^{-1} \begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix} \right\}, \quad (2.5)$$

with respect to  $\beta$  and  $b$ , which ultimately requires minimizing the criterion

$$(y - X\beta - Zb)' \Sigma^{-1} (y - X\beta - Zb) + b' D^{-1} b. \quad (2.6)$$

This leads to the mixed model equations

$$\begin{pmatrix} X' \Sigma^{-1} X & X' \Sigma^{-1} Z \\ Z' \Sigma^{-1} X & Z' \Sigma^{-1} Z + D^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X' \Sigma^{-1} y \\ Z' \Sigma^{-1} y \end{pmatrix}. \quad (2.7)$$

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  $\Sigma$  are known.

? points out that although ? initially referred to the estimates  $\hat{\beta}$  and  $\hat{b}$  from (2.7) as “joint maximum likelihood estimates”, ? later advised that these estimates should not be referred to as “maximum likelihood” as the function being maximized in (2.6) is a joint density rather than a likelihood function. ? remarks that it is clear that Henderson used joint estimation for computational purposes, without recognizing the theoretical implications.

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

However these estimates are known to be biased ‘downwards’ (i.e. underestimated), because of the assumption that the fixed estimates are known, rather than being estimated from the data. ? 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.

### 2.1.3 Likelihood and estimation

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

Likelihood functions provide the basis for two important statistical concepts that shall be further referred to; the likelihood ratio test and the Akaike information criterion. The likelihood function,  $L(\theta)$ , is a fundamental concept in statistical inference. It indicates how likely a particular population is to produce an observed sample. The set of values that maximize the likelihood function are considered to be optimal, and

are used as the estimates of the parameters. For computational ease, it is common to use the logarithm of the likelihood function, known simply as the log-likelihood ( $\ell(\theta)$ ).

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

### **Likelihood-based tools**

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

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

### **2.1.4 Algorithms**

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

The most common iterative algorithms used for the optimization problem in the context of LMEs are the EM algorithm, fisher scoring algorithm and NR algorithm, which ? commends as the preferred method. Parameter of the mixed model can

be estimated using either ML or REML, while the AIC and the BIC can be used as measures of “goodness of fit” for particular models, where smaller values are considered preferable.

## Restricted Likelihood Estimation

Restricted maximum likelihood (REML) is an alternative methods of computing parameter estimated, developed by ? and ? to provide unbiased estimates of variance and covariance parameters. The REML approach does not base estimates on a maximum likelihood fit of all the information, but instead uses a likelihood function derived from a data set, transformed to remove the irrelevant influences (?).

REML obtains estimates of the fixed effects using non-likelihoodlike methods, such as ordinary least squares or generalized least squares, and then using these estimates it maximizes the likelihood of the residuals (subtracting off the fixed effects) to obtain estimates of the variance parameters. In most software packages REML is the default algorithm used to compute coefficients for the predictor variables. REML estimation reduces the bias in the variance component, and also handles high correlations more effectively, and is less sensitive to outliers than ML.

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

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

Restricted maximum likelihood is often preferred to maximum likelihood because REML estimation reduces the bias in the variance component by taking into account the loss of degrees of freedom that results from estimating the fixed effects in  $\beta$ . Re-

stricted 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.

### 2.1.5 Estimation of Fixed and Random Effects

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

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

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

## 2.2 LME models in method comparison studies

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

In part this is due to the increased profile of LME models, and furthermore the availability of capable software. Additionally a great understanding of residual analysis and influence analysis for LME models has been achieved thanks to authors such as ?, ?, ? ?, amongst others.

Due to the prevalence of modern statistical software, ? advocates the adoption of computer based approaches to method comparison studies, allowing the use of LME models that would not have been feasible otherwise. These authors remark that modern statistical computation, such as that used for LME models, greatly improve the efficiency of calculation compared to previous ‘by-hand’ approaches, as advocated in ?, describing them as tedious, unnecessary and ‘outdated’. Rather than using the ‘by hand’ methods, estimates for required LME parameters can be read directly from program output. Furthermore, using computer approaches removes associated constraints, such as the need for the design to be perfectly balanced.

? 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, ? 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.



? views the uses of linear mixed effects models as an expansion on the Bland-Altman methodology, rather than as a replacement.

Their focus is to explain lack of agreement by means of additional covariates outside the scope of the traditional method comparison problem, which extends beyond the conventional method comparison study question. The data used for their examples is unavailable for independent use. Therefore, for the sake of consistency, a data set will be simulated based on the Blood Data that will allow for extra variables, and an exploration shall be provided in the appendices.

### 2.2.1 Agreement Criteria

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

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

Further to this, ? demonstrates an suite of tests that can be used to determine how well two methods of measurement, in the presence of repeated measures, agree with

each other.

- No Significant inter-method bias
- No difference in the between-subject variabilities of the two methods
- No difference in the within-subject variabilities of the two methods

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

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

These tests are complemented by the ability to consider the inter-method bias and the overall correlation coefficient. Two methods can be considered to be in agreement if criteria based upon these methodologies are met. Additionally Roy makes reference to the overall correlation coefficient of the two methods, which is determinable from variance estimates. Other important aspects of the method comparison study are consequent. The limits of agreement are computed using the results of the reference model.

Differences in between-subject variabilities of the two methods arise when one method is yielding average response levels for individuals that 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 that 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.

### 2.2.2 Fitting LME Models to Method Comparison Data

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

### 2.2.3 Roy's Approach

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

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

Importantly ? 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.

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

based on ?, which in turn follows on from ?.

? 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. 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.

The well-known “Limits of Agreement”, as developed by ? are not referred to directly, but are easily computable using the framework proposed by ?. Further discussion will be provided in due course.

#### **2.2.4 Replicate measurements in Roy’s paper**

Measurements taken in quick succession by the same observer using the same instrument on the same subject can be considered true replicates. ? notes that some measurements may not be ‘true’ replicates.

Roy’s methodology assumes the use of “true replicates”. However data may not be collected in this way. In such cases, the correlation matrix on the replicates may require a different structure, such as the autoregressive order one  $AR(1)$  structure. However determining MLEs with such a structure would be computational intense, if possible at all.

? takes its definition of replicate measurement: two or more measurements on the same item taken under identical conditions.

#### **2.2.5 Specifying the Models**

Roy proposes a series of three tests on the variance components of an LME model. For these tests, four candidate models are constructed.

Using Roy’s method, four candidate models are constructed, each differing by con-

straints applied to the variance covariance matrices. In addition to computing the inter-method bias, three significance tests are carried out on the respective formulations to make a judgement on whether or not two methods are in agreement.

The difference in the models are specifically in how the  $D$  and  $\Sigma$  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.

? considers four independent hypothesis tests. The first test allows of the comparison the begin-subject variability of two methods. Similarly, the second test assesses the within-subject variability of two methods. A third test is a test that compares the overall variability of the two methods.

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

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

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

## 2.2.6 Model Specification for Roy's Hypotheses Tests

Response for  $i$ th subject can be written as

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

- $\beta_1$  and  $\beta_2$  are fixed effects corresponding to both methods. ( $\beta_0$  is the intercept.)
- $b_{1i}$  and  $b_{2i}$  are random effects corresponding to both methods.

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

## 2.2.7 LME Model Specification

Let  $y_{mir}$  denote the  $r$ th replicate measurement on the  $i$ th item by the  $m$ th method, where  $m = 1, 2$ ,  $i = 1, \dots, N$ , and  $r = 1, \dots, n_i$ . 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

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \quad (2.8)$$

Here  $\beta_0$  and  $\beta_m$  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  $E(b_{mi}) = 0$  with  $\text{Var}(b_{mi}) = g_m^2$  and  $\text{Cov}(b_{mi}, b_{m'i}) = d_{12}$ . The random error term for each response is denoted  $\epsilon_{mir}$  having  $E(\epsilon_{mir}) = 0$ ,  $\text{Var}(\epsilon_{mir}) = \sigma_m^2$ ,  $\text{Cov}(b_{mir}, b_{m'ir}) = \sigma_{12}$ ,  $\text{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$  and  $\text{Cov}(\epsilon_{mir}, \epsilon_{m'ir'}) = 0$ .

When two methods of measurement are in agreement, there is no significant differences between  $\beta_1$  and  $\beta_2$ ,  $g_1^2$  and  $g_2^2$ , and  $\sigma_1^2$  and  $\sigma_2^2$ . Here  $\beta_0$  and  $\beta_m$  are fixed-effect terms representing, respectively, a model intercept and an overall effect for method  $m$ .

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

The random error term for each response is denoted  $\epsilon_{mir}$  having  $E(\epsilon_{mir}) = 0$ ,  $\text{Var}(\epsilon_{mir}) = \sigma_m^2$ ,  $\text{Cov}(\epsilon_{1ir}, \epsilon_{2ir}) = \sigma_{12}$ ,  $\text{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$  and  $\text{Cov}(\epsilon_{1ir}, \epsilon_{2ir'}) = 0$ . Two methods of measurement are in complete agreement if the null hypotheses  $H_1: \alpha_1 = \alpha_2$  and  $H_2: \sigma_1^2 = \sigma_2^2$  and  $H_3: g_1^2 = g_2^2$  hold simultaneously. ? 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.

### 2.2.8 Test for inter-method bias

Firstly, a practitioner would investigate whether a significant inter-method bias is present between the methods. This bias is specified as a fixed effect in the LME model. For a practitioner who has a reasonable level of competency in R and undergraduate statistics (in particular simple linear regression model) this is a straight-forward procedure.

The presence of an inter-method bias is the source of disagreement between two methods of measurement that is most easily identified. As the first in a series of hypothesis tests, ? 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

$$H_1 : \mu_1 = \mu_2,$$

$$K_1 : \mu_1 \neq \mu_2.$$

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

### 2.2.9 Roy's hypothesis tests : Roy's variability tests

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

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

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

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

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

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

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.

### 2.2.10 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  $2 \times 2$  matrices.



A discussion of the various structures a variance-covariance matrix can be specified under is required before progressing. The following structures are relevant: the identity structure, the compound symmetric structure and the symmetric structure.

The differences in the models are specifically in how the the  $D$  and  $\Lambda$  matrices are constructed, using either an unstructured form or a compound symmetry form. To illustrate these differences, consider a generic matrix  $A$ ,

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

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

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

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

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  $\psi^2$ ) for all 3 of the time points and a single covariance (represented by  $\psi_{ij}$ ) for each of the pairs of trials.

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

### 2.2.11 Variability Tests

? further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal. An examination of this topic is useful because a method for computing Limits of Agreement follows from here.

Variability tests proposed by ? affords the opportunity to expand upon Carstensen's approach. Three tests of hypothesis are provided, appropriate for evaluating the agreement between the two methods of measurement under this sampling scheme.

#### Model 1

The first candidate model is compared to each of the three other models successively. It is the alternative model in each of the three tests, with the other three models acting as the respective null models. The models are compared using the likelihood ratio test, a general method for comparing nested models fitted by ML (?).

The first model acts as an alternative hypothesis to be compared against each of three other models, acting as null hypothesis models, successively. The models are compared using the likelihood ratio test. Likelihood ratio tests are a class of tests based on the comparison of the values of the likelihood functions of two candidate models. The first test allows of the comparison the begin-subject variability of two methods.

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

The first test allows of the comparison the begin-subject variability of two methods. Similarly, the second test assesses the within-subject variability of two methods. A third test is a test that compares the overall variability of the two methods.

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

? proposes a suite of hypothesis tests for assessing the agreement of two methods of measurement, when replicate measurements are obtained for each item, using a LME approach. The tests are implemented by fitting a specific LME model, and three variations thereof, to the data. These three variant models introduce equality constraints that act null hypothesis cases. Two methods of measurement are in complete agreement if the null hypotheses  $H_1: \alpha_1 = \alpha_2$  and  $H_2: \sigma_1^2 = \sigma_2^2$  and  $H_3: d_1^2 = d_2^2$  hold simultaneously. ? 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.

### Variability test 1

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

$$H_0: d_1 = d_2$$

$$H_A: d_1 \neq d_2$$

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

### Variability test 2

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

$$H_0 : \sigma_1 = \sigma_2$$

$$H_A : \sigma_1 \neq \sigma_2$$

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

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

### Variability test 3

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

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

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

$$H_0 : \omega_1 = \omega_2$$

$$H_A : \omega_1 \neq \omega_2$$

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

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

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \quad (2.9)$$

Overall variability between the two methods ( $\Omega$ ) is sum of between-subject ( $D$ ) and within-subject variability ( $\Sigma$ ),  $\Omega$  denotes the overall variability as Block -  $\Omega_i$ . The overall variation for methods 1 and 2 are given by

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

### 2.2.12 Computing Limits of Agreement

The variance of case-wise difference in measurements can be determined from Block- $\Omega_i$ . 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 -  $\Omega_i$  matrix. Lack of agreement can arise if there is a disagreement in overall variabilities.

### 2.2.13 Formal testing for covariances (Off-Diagonal Components in Roy’s Model)

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

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

$\sigma_x^2$  and  $\sigma_y^2$  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.

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} \text{ vs } \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$$

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

$$H_5 : \sigma_{12} = 0$$

$$K_5 : \sigma_{12} \neq 0$$

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.

## 2.2.14 Correlation coefficient

Roy's tests are complemented by the ability to the overall correlation coefficient of the two methods, which is determinable from variance estimates. Two methods can be considered to be in agreement if criteria based upon these tests are met. Inference for inter-method bias follows from well-established methods and, as such, will only be noted when describing examples.

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

? 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.

? 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  $\rho_{xy}$  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.

### 2.2.15 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.

### 2.2.16 Roy's methodology for single measurements

Roy's methodology is not suitable for the case of single measurements because it follows from the decomposition for the covariance matrix of the response vector  $y_i$ , as presented in ?. 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.

## 2.3 (good)Limits of agreement in LME models

Limits of agreement are used extensively for assessing agreement, because they are intuitive and easy to use. Necessarily their prevalence in literature has meant that they are now the best known measurement for agreement, and therefore any newer methodology would benefit by making reference to them.

? uses LME models to determine the limits of agreement. Between-subject variation for method  $m$  is given by  $d_m^2$  and within-subject variation is given by  $\lambda_m^2$ . ? remarks



that for two methods  $A$  and  $B$ , separate values of  $d_A^2$  and  $d_B^2$  cannot be estimated, only their average. Hence the assumption that  $d_x = d_y = d$  is necessary. The between-subject variability  $\mathbf{D}$  and within-subject variability  $\mathbf{\Lambda}$  can be presented in matrix form,

$$\mathbf{D} = \begin{pmatrix} d_A^2 & 0 \\ 0 & d_B^2 \end{pmatrix} = \begin{pmatrix} d^2 & 0 \\ 0 & d^2 \end{pmatrix}, \quad \mathbf{\Lambda} = \begin{pmatrix} \lambda_A^2 & 0 \\ 0 & \lambda_B^2 \end{pmatrix}.$$

The variance for method  $m$  is  $d_m^2 + \lambda_m^2$ . Limits of agreement are determined using the standard deviation of the case-wise differences between the sets of measurements by two methods  $A$  and  $B$ , given by

$$\text{var}(y_A - y_B) = 2d^2 + \lambda_A^2 + \lambda_B^2. \quad (2.10)$$

Importantly the covariance terms in both variability matrices are zero, and no covariance component is present.

? presents a data set ‘fat’, which is a comparison of measurements of subcutaneous fat by two observers at the Steno Diabetes Center, Copenhagen. Measurements are in millimeters (mm). Each person is measured three times by each observer. The observations are considered to be ‘true’ replicates.

A linear mixed effects model is formulated, and implementation through several software packages is demonstrated. All of the necessary terms are presented in the computer output. The limits of agreement are therefore,

$$0.0449 \pm 1.96 \times \sqrt{2 \times 0.0596^2 + 0.0772^2 + 0.0724^2} = (-0.220, 0.309). \quad (2.11)$$

? has demonstrated a methodology whereby  $d_A^2$  and  $d_B^2$  can be estimated separately. Also covariance terms are present in both  $\mathbf{D}$  and  $\mathbf{\Lambda}$ . Using Roy’s methodology, the variance of the differences is

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_A^2 + \lambda_B^2 - 2(d_{AB} + \lambda_{AB}) \quad (2.12)$$

All of these terms are given or determinable in computer output. The limits of agreement can therefore be evaluated using

$$\bar{y}_A - \bar{y}_B \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \quad (2.13)$$

For Carstensen’s ‘fat’ data, the limits of agreement computed using Roy’s method are consistent with the estimates given by ?;  $0.044884 \pm 1.96 \times 0.1373979 = (-0.224, 0.314)$ .

### 2.3.1 (good)Linked replicates

? proposes the addition of an random effects term to their model when the replicates are linked. This term is used to describe the ‘item by replicate’ interaction, which is independent of the methods. This interaction is a source of variability independent of the methods. Therefore failure to account for it will result in variability being wrongly attributed to the methods.

? introduces a second data set; the oximetry study. This study done at the Royal Childrens Hospital in Melbourne to assess the agreement between co-oximetry and pulse oximetry in small babies.

In most cases, measurements were taken by both method at three different times. In some cases there are either one or two pairs of measurements, hence the data is unbalanced. ? describes many of the children as being very sick, and with very low oxygen saturations levels. Therefore it must be assumed that a biological change can occur in interim periods, and measurements are not true replicates.

? demonstrate the necessity of accounting for linked replicated by comparing the limits of agreement from the ‘oximetry’ data set using a model with the additional term, and one without. When the interaction is accounted for the limits of agreement are  $(-9.62, 14.56)$ . When the interaction is not accounted for, the limits of agreement are  $(-11.88, 16.83)$ . It is shown that the failure to include this additional term results in an over-estimation of the standard deviations of differences.

Limits of agreement are determined using Roy’s methodology, without adding any additional terms, are found to be consistent with the ‘interaction’ model;  $(-9.562, 14.504)$ . Roy’s methodology assumes that replicates are linked. However, following Carstensen’s example, an addition interaction term is added to the implementation of Roy’s model to assess the effect, the limits of agreement estimates do not change. However there

is a conspicuous difference in within-subject matrices of Roy’s model and the modified model (denoted 1 and 2 respectively);

$$\hat{\mathbf{\Lambda}}_1 = \begin{pmatrix} 16.61 & 11.67 \\ 11.67 & 27.65 \end{pmatrix} \quad \hat{\mathbf{\Lambda}}_2 = \begin{pmatrix} 7.55 & 2.60 \\ 2.60 & 18.59 \end{pmatrix}. \quad (2.14)$$

(The variance of the additional random effect in model 2 is 3.01.)

? 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. Two candidate models can said to be equally good if there is a difference of less than 2 in their AIC values.

The Akaike information criterion ( $AIC$ ) for both models are  $AIC_1 = 2304.226$  and  $AIC_2 = 2306.226$ , indicating little difference in models. The AIC values for the Carstensen ‘unlinked’ and ‘linked’ models are 1994.66 and 1955.48 respectively, indicating an improvement by adding the interaction term.

The  $\hat{\mathbf{\Lambda}}$  matrices are informative as to the difference between Carstensen’s unlinked and linked models. For the oximetry data, the covariance terms (given above as 11.67 and 2.6 respectively) are of similar magnitudes to the variance terms. Conversely for the ‘fat’ data the covariance term ( $-0.00032$ ) is negligible. When the interaction term is added to the model, the covariance term remains negligible. (For the ‘fat’ data, the difference in AIC values is also approximately 2).

To conclude, Carstensen’s models provided a rigorous way to determine limits of agreement, but don’t provide for the computation of  $\hat{\mathbf{D}}$  and  $\hat{\mathbf{\Lambda}}$ . Therefore the test’s proposed by ? can not be implemented. Conversely, accurate limits of agreement as determined by Carstensen’s model may also be found using Roy’s method. Addition of the interaction term erodes the capability of Roy’s methodology to compare candidate models, and therefore shall not be adopted.

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

## 2.4 (good) Conclusion

? and ? highlight the need for method comparison methodologies suitable for use in the presence of replicate measurements. ? 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.

# Bibliography

## 2.5 Likelihood Ratio Tests

Likelihood ratio tests (LRTs) are a family of tests used to compare the value of likelihood functions for two models, whose respective formulations define a hypothesis to be tested (i.e. the nested and reference model). LRTs can be used to test hypotheses about covariance parameters or fixed effects parameters in the context of LMEs. Each of these three test shall be examined in more detail shortly.

The test statistic for each of the three hypothesis tests is the difference of the  $M2LL$  for each pair of models. The probability distribution of the test statistic is approximated by the  $\chi^2$  distribution with  $(\nu_1 - \nu_2)$  degrees of freedom, where  $\nu_1$  and  $\nu_2$  are the degrees of freedom of models 1 and 2 respectively. If the  $p$ -value in each of the respective tests exceed as significance level chosen by the analyst, then the null model must be rejected. The significance of the likelihood ratio test can be found by comparing the likelihood ratio to the  $\chi^2$  distribution, with the appropriate degrees of freedom.

$$\nu = [\text{LRT df under } H_0 \text{ model}] - [\text{LRT df under } H_A \text{ model}] \quad (2.15)$$

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

$$-2 \ln \Lambda_d = [M2LL \text{ under } H_0 \text{ model}] - [M2LL \text{ under } H_A \text{ model}] \quad (2.16)$$

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

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

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

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

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

and the maximum likelihood estimate is the solution to the score equation  $S(\theta) = 0$ . The Fisher information  $I(\theta)$ , which is defined as

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

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

### 2.5.1 Statistical Assumptions for Likelihood Ratio Tests

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

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

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 likelihood), but only if the fixed terms in the two models are invariant, and

both models have been fit by REML. Otherwise, the argument: `method="ML"` must be employed (ML = maximum likelihood).

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

### 2.5.2 Nesting: Model Selection Using Likelihood Ratio Tests

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

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

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

### 2.5.3 Relevance of Estimation Methods

The problem with REML for model building is that the “likelihoods” obtained for different fixed effects are not comparable. Hence it is not valid to compare models with different fixed effects using a likelihood ratio test or AIC when REML is used to estimate the model. Therefore models derived using ML must be used instead.

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

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

For both REML and ML estimates, the nominal  $p$ -values for the LRT statistics under a  $\chi^2$  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.

#### 2.5.4 Akaike Information Critierion

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

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

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

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



# Bibliography