

Contents

1	Introduction	6
1.1	LME models in method comparison studies	6
1.2	LRTs	8
1.3	Difference Between Approaches	8
1.4	Carstensen's Model	9
1.5	Two Way ANOVA	9
1.5.1	Standard Two Way ANOVA	11
1.5.2	BXC2004 Model	12
1.5.3	Statement of the LME model	12
1.6	Agreement Criteria	13
2	Model Specification	16
2.1	Model Formula	18
2.2	Formulation of the response vector	20
2.3	Decomposition of the response covariance matrix	20
2.4	Sampling Scheme : Linked and Unlinked Replicates	22
2.5	Replicate measurements	22
3	Introduction to Roy's Procedure	23
3.1	Introduction to Roy's methodology	24
3.2	Model Set Up	25
3.3	Criteria	25

3.4	Test for inter-method bias	26
3.5	Variability Tests	27
3.6	Roy's Candidate Models	27
3.7	Likelihood Ratio Tests	29
3.7.1	Likelihood Ratio Tests	30
3.7.2	Likelihood Ratio Tests	31
3.8	Model Selection Using Likelihood Ratio Tests	32
3.8.1	Likelihood Ratio Tests	33
3.9	Likelihood Ratio Tests	34
3.9.1	Pinheiro Bates	34
3.9.2	Empirical p-values of LRT tests	34
3.9.3	Other material	35
3.9.4	Nested and Reference Models	36
3.9.5	LRTs for covariance parameters	36
3.9.6	Likelihood Ratio Tests	37
3.9.7	Likelihood Ratio Tests	38
3.9.8	Relevance of Estimation Methods	38
3.9.9	Other material	41
3.9.10	Nested and Reference Models	41
3.9.11	Test for inter-method bias	42
3.10	Likelihood Ratio Tests	43
3.11	Roy's LME methodology for assessing agreement	43
4	Model Specification	45
4.1	Model Specification for Roy's Hypotheses Tests	45
4.2	G Component	48
4.3	R Component	48
4.4	Hamlett	51
4.5	For Expository Purposes	52

4.6	Kroneckor	53
4.7	Overall Variability	54
5	Roy's hypothesis tests	56
5.1	Hypothesis Testing	56
5.2	Roy's hypothesis tests	57
5.3	Roy's variability tests	58
5.4	Variability test 1	59
5.5	Variability test 2	60
5.6	Variability test 3	61
5.7	Omnibus testing	61
5.8	Implementation	62
5.9	Formal testing for covariances	62
5.9.1	Variance Covariance Matrices	62
5.10	VC structures	64
6	LOAS	66
6.1	Calculation of limits of agreement	66
6.2	00-Limits of Agreement in LME models	68
6.3	Carstensen et al	68
6.4	Computing LoAs from LME models	69
6.5	Carstensen's Limits of Agreement	69
6.6	Carstensen's Model	70
6.7	BXC2008 presents - Carstensen's Limits of agreement	73
6.8	Limits of Agreement in LME models	74
6.9	Limits of Agreement in LME models	75
6.10	BXC2008 uses - Limits of Agreement in LME models	75
6.11	Carstensen's LOAs	76
6.12	Interaction Terms in Model	76
6.13	Computation of limits of agreement	77

6.14	BXC - Model Terms	77
6.15	Computation of limits of agreement under Roy's model	78
6.16	LOAs with Roy	79
6.17	Differences Between Models	79
6.18	Differences	81
6.19	Relevance of Roy's Methodology	82
6.20	Difference Variance further to Carstensen	82
6.21	Assumptions on Variability	83
7	Other Material	85
7.1	Extension of Roy's methodology	85
7.1.1	Correlation	86
7.1.2	Correlation terms	87
7.2	Hamlett and Lam	88
7.3	LaiShiao	88
7.4	LaiShiao	89
7.5	Lai Shiao	90
7.6	Lai Shiao	92
7.7	Lai Shiao	94
7.8	Demonstration of Roy's testing	97
7.9	Worked Eamples	99
7.9.1	LikelihoodRatio Tests	99
7.9.2	Reference Model (Ref.Fit)	102
7.9.3	Nested Model (Between-Item Variability)	102
7.10	Testing Procedures	102
7.10.1	Roy's Reference Model	104
7.10.2	Nested Model (Between-Item Variability)	104
7.11	Worked Eamples	106
7.12	Examples: LoAs for Carstensen's data	106

7.12.1	Daibetes Example	107
7.12.2	Oximetry Data	107
7.12.3	Linked replicates	111
7.12.4	Limits of agreement for Carstensen's data	113
7.12.5	Fat Data Examples: LoAs for Carstensen's data	113
7.12.6	RV-IV	114
7.12.7	Limits of agreement for Oximetry	114
7.12.8	Classical Model	115
7.13	Classical model for single measurements	115
7.13.1	Sampling	118
7.13.2	Remarks on the Multivariate Normal Distribution	119
7.14	Implementation in R	122
7.15	Conclusion	125
	Bibliography	125

Chapter 1

Introduction

1.1 LME models in method comparison studies

Barnhart et al. (2007) describes the sources of disagreement in a method comparison study problem as differing population means, different between-subject variances, different within-subject variances between two methods and poor correlation between measurements of two methods. Further to this, Roy (2009b) states three criteria for two methods to be considered in agreement. Firstly that there be no significant bias. Second that there is no difference in the between-subject variabilities, and lastly that there is no significant difference in the within-subject variabilities. Roy further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal.

Roy (2009b) further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal.

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

Lai and Shiao (2005) views the uses of linear mixed effects models as an expansion

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

Lai and Shiao (2005) extends the usual method comparison study question. It correctly identifies LME models as a methodology that can used to make such questions tractable. The Data Set used in 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. Carstensen et al. (2008) remarks that modern statistical computation, such as that used for LME models, greatly improve the efficiency of calculation compared to previous ‘by-hand’ methods. Additionally a great understanding of residual analysis and influence analysis for LME models has been achieved thanks to authors such as ?, ?, Cook (1986) West et al. (2007), amongst others. In this chapter various LME approaches to method comparison studies shall be examined.

Due to the prevalence of modern statistical software, Carstensen et al. (2008) advocates the adoption of computer based approaches, such as LME models, to method comparison studies. Carstensen et al. (2008) remarks upon ‘by-hand’ approaches advocated in Bland and Altman (1999) discouragingly, describing them as tedious, unnecessary and ‘outdated’. Rather than using the ‘by hand’ methods, estimates for required LME parameters can be read directly from program output. Furthermore, using computer approaches removes constraints associated with ‘by-hand’ approaches, such as the need for the design to be perfectly balanced.

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

1.2 LRTs

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

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

1.3 Difference Between Approaches

Carstensen presents two models. One for the case where the replicates, and a second for when they are linked.

Carstensen's model does not take into account either between-item or within-item covariance between methods.

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

$$\begin{pmatrix} \omega_2^1 & 0 \\ 0 & \omega_2^2 \end{pmatrix} = \begin{pmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

1.4 Carstensen's Model

Carstensen (2004) presents a model to describe the relationship between a value of measurement and its real value. The non-replicate case is considered first, as it is the context of the Bland Altman plots. This model assumes that inter-method bias is the only difference between the two methods.

A measurement y_{mi} by method m on individual i is formulated as follows;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2) \quad (1.1)$$

The differences are expressed as $d_i = y_{1i} - y_{2i}$. For the replicate case, an interaction term c is added to the model, with an associated variance component. All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \quad (1.2)$$

Of particular importance is terms of the model, a true value for item i (μ_i). The fixed effect of Roy's model comprise of an intercept term and fixed effect terms for both methods, with no reference to the true value of any individual item. A distinction can be made between the two models: Roy's model is a standard LME model, whereas Carstensen's model is a more complex additive model.

1.5 Two Way ANOVA

Carstensen et al. (2008) develop their model from a standard two-way analysis of variance model, reformulated for the case of replicate measurements, with random effects terms specified as appropriate. Their model can be written as

$$y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + \varepsilon_{mir}. \quad (1.3)$$

The fixed effects α_m and μ_i represent the intercept for method m and the ‘true value’ for item i respectively. The random-effect terms comprise an item-by-replicate interaction term $a_{ir} \sim \mathcal{N}(0, \varsigma^2)$, a method-by-item interaction term $c_{mi} \sim \mathcal{N}(0, \tau_m^2)$, and model error terms $\varepsilon_{mir} \sim \mathcal{N}(0, \varphi_m^2)$. All random-effect terms are assumed to be independent. For the case when replicate measurements are assumed to be exchangeable for item i , a_{ir} can be removed. The model expressed in (2) describes measurements by m methods, where $m = \{1, 2, 3 \dots\}$. Based on the model expressed in (2), Carstensen et al. (2008) compute the limits of agreement as

$$\alpha_1 - \alpha_2 \pm 2\sqrt{\tau_1^2 + \tau_2^2 + \varphi_1^2 + \varphi_2^2}$$

Carstensen et al. (2008) notes that, for $m = 2$, separate estimates of τ_m^2 can not be obtained. To overcome this, the assumption of equality, i.e. $\tau_1^2 = \tau_2^2$ is required.

There is a substantial difference in the number of fixed parameters used by the respective models; the model in (2.3) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items N , whereas the model in (1.4) requires $N + 2$ fixed effects.

Allocating fixed effects to each item i by (1.4) accords with earlier work on comparing methods of measurement, such as Grubbs (1948). However allocation of fixed effects in ANOVA models suggests that the group of items is itself of particular interest, rather than as a representative sample used of the overall population. However this approach seems contrary to the purpose of LOAs as a prediction interval for a population of items. Conversely, Roy (2009a) uses a more intuitive approach, treating the observations as a random sample population, and allocating random effects accordingly.

1.5.1 Standard Two Way ANOVA

Carstensen et al. (2008) develop their model from a standard two-way analysis of variance model, reformulated for the case of replicate measurements, with random effects terms specified as appropriate. Their model describing y_{mir} , again the r th replicate measurement on the i th item by the m th method ($m = 1, 2$, $i = 1, \dots, N$, and $r = 1, \dots, n$), can be written as

$$y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + \epsilon_{mir}. \quad (1.4)$$

The fixed effects α_m and μ_i represent the intercept for method m and the ‘true value’ for item i respectively. The random-effect terms comprise an item-by-replicate interaction term $a_{ir} \sim \mathcal{N}(0, \varsigma^2)$, a method-by-item interaction term $c_{mi} \sim \mathcal{N}(0, \tau_m^2)$, and model error terms $\epsilon \sim \mathcal{N}(0, \varphi_m^2)$. All random-effect terms are assumed to be independent. For the case when replicate measurements are assumed to be exchangeable for item i , a_{ir} can be removed.

There is a substantial difference in the number of fixed parameters used by the respective models. For the model in (2.3) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items N . In contrast, the model described by (1.4) requires $N + 2$ fixed effects for N items. The inclusion of fixed effects to account for the ‘true value’ of each item greatly increases the level of model complexity.

When only two methods are compared, Carstensen et al. (2008) notes that separate estimates of τ_m^2 can not be obtained due to the model over-specification. To overcome this, the assumption of equality, i.e. $\tau_1^2 = \tau_2^2$, is required.

1.5.2 BXC2004 Model

Carstensen (2004) presents a model to describe the relationship between a value of measurement and its real value. The non-replicate case is considered first, as it is the context of the Bland Altman plots. This model assumes that inter-method bias is the only difference between the two methods.

A measurement y_{mi} by method m on individual i is formulated as follows;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2) \quad (1.5)$$

The differences are expressed as $d_i = y_{1i} - y_{2i}$. For the replicate case, an interaction term c is added to the model, with an associated variance component. All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \quad (1.6)$$

Of particular importance is terms of the model, a true value for item i (μ_i). The fixed effect of Roy's model comprise of an intercept term and fixed effect terms for both methods, with no reference to the true value of any individual item. A distinction can be made between the two models: Roy's model is a standard LME model, whereas Carstensen's model is a more complex additive model.

1.5.3 Statement of the LME model

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

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

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

As such, linear mixed effects models treat the error for fixed effects differently than the error for random effects. A linear mixed effects model is a linear model that combined fixed and random effect terms formulated by Laird and Ware (1982) as follows;

$$Y_i = X_i\beta + Z_ib_i + \epsilon_i$$

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

1.6 Agreement Criteria

Roy (2009b) 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. (An item would commonly be a patient).

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

Firstly, there is no inter-method bias between the two methods, i.e. there is no persistent tendency for one method to give higher values than the other.

Secondly, both methods of measurement have the same within-subject variability. In such a case the variance of the replicate measurements would consistent for both methods. Lastly, the methods have equal between-subject variability. Put simply, for the mean measurements for each case, the variances of the mean measurements from both methods are equal.

Lack of agreement can arise if there is a disagreement in overall variabilities. This may be due to the disagreement in either between-item variabilities or within-item variabilities, or both. Roy (2009b) allows for a formal test of each.

Roy (2009b) sets out three criteria for two methods to be considered in agreement. Firstly that there be no significant bias. Second that there is no difference in the between-subject variabilities, and lastly that there is no significant difference in the within-subject variabilities. Roy further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal.

Two methods of measurement are in complete agreement if the null hypotheses $H_1: \alpha_1 = \alpha_2$ and $H_2: \sigma_1^2 = \sigma_2^2$ and $H_3: g_1^2 = g_2^2$ hold simultaneously. Roy (2009b) uses a Bonferroni correction to control the familywise error rate for tests of $\{H_1, H_2, H_3\}$ and account for difficulties arising due to multiple testing. Roy also integrates H_2 and H_3 into a single testable hypothesis $H_4: \omega_1^2 = \omega_2^2$, where $\omega_m^2 = \sigma_m^2 + g_m^2$ represent the overall variability of method m . Disagreement in overall variability may be caused by different between-item variabilities, by different within-item variabilities, or by both. If the exact cause of disagreement between the two methods is not of interest, then the overall variability test H_4 is an alternative to testing H_2 and H_3 separately.

(Work this in) Roy’s method considers two methods to be in agreement if three conditions are met.

- no significant bias, i.e. the difference between the two mean readings is not “statistically significant”,
- high overall correlation coefficient,
- the agreement between the two methods by testing their repeatability coefficients.

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

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

Chapter 2

Model Specification

Model Terms (Roy 2009)

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

Model Terms (Roy 2009)

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

- Let the number of replicate measurements on each item i for both methods be n_i , hence $2 \times n_i$ responses. However, it is assumed that there may be a different number of replicates made for different items. Let the maximum number of replicates be p . An item will have up to $2p$ measurements, i.e. $\max(n_i) = 2p$.
- Later on \mathbf{X}_i will be reduced to a 2×1 matrix, to allow estimation of terms. This is due to a shortage of rank. The fixed effects vector can be modified accordingly.
- \mathbf{Z}_i is the $2n_i \times 2$ model matrix for the random effects for measurement methods on item i .
- \mathbf{b}_i is the 2×1 vector of random-effect coefficients on item i , one for each method.
- $\boldsymbol{\epsilon}$ is the $2n_i \times 1$ vector of residuals for measurements on item i .
- \mathbf{G} is the 2×2 covariance matrix for the random effects.
- \mathbf{R}_i is the $2n_i \times 2n_i$ covariance matrix for the residuals on item i .
- The expected value is given as $E(\mathbf{y}_i) = \mathbf{X}_i\boldsymbol{\beta}$. (Hamlett et al., 2004)
- The variance of the response vector is given by $\text{Var}(\mathbf{y}_i) = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$ (Hamlett et al., 2004).

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

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \quad (2.1)$$

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

$$\mathbf{b}_i = \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} \quad (2.2)$$

- \mathbf{V} represents the correlation matrix of the replicated measurements on a given method. $\mathbf{\Sigma}$ is the within-subject VC matrix.
- \mathbf{V} and $\mathbf{\Sigma}$ are positive definite matrices. The dimensions of \mathbf{V} and $\mathbf{\Sigma}$ are $3 \times 3 (= p \times p)$ and $2 \times 2 (= k \times k)$.
- It is assumed that \mathbf{V} is the same for both methods and $\mathbf{\Sigma}$ is the same for all replications.
- $\mathbf{V} \otimes \mathbf{\Sigma}$ creates a $6 \times 6 (= kp \times kp)$ matrix. \mathbf{R}_i is a sub-matrix of this.

2.1 Model Formula

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

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \quad (2.3)$$

Here β_0 and β_m are fixed-effect terms representing, respectively, a model intercept and an overall effect for method m . The b_{1i} and b_{2i} terms represent random effect parameters corresponding to the two methods, having $E(b_{mi}) = 0$ with $\text{Var}(b_{mi}) = g_m^2$ and $\text{Cov}(b_{mi}, b_{m'i}) = g_{12}$. The random error term for each response is denoted ϵ_{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 β_1 and β_2 , g_1^2 and g_2^2 , and σ_1^2 and σ_2^2 . Here β_0 and β_m are fixed-effect terms representing, respectively, a model intercept and an overall effect

for method m . The model can be reparameterized by gathering the β terms together into (fixed effect) intercept terms $\alpha_m = \beta_0 + \beta_m$. The b_{1i} and b_{2i} terms are correlated random effect parameters having $E(b_{mi}) = 0$ with $\text{Var}(b_{mi}) = g_m^2$ and $\text{Cov}(b_{1i}, b_{2i}) = g_{12}$. The random error term for each response is denoted ϵ_{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. Roy (2009a) 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. Roy also integrates H_2 and H_3 into a single testable hypothesis $H_4: \omega_1^2 = \omega_2^2$, where $\omega_m^2 = \sigma_m^2 + g_m^2$ represent the overall variability of method m . Disagreement in overall variability may be caused by different between-item variabilities, by different within-item variabilities, or by both. If the exact cause of disagreement between the two methods is not of interest, then the overall variability test H_4 is an alternative to testing H_2 and H_3 separately.

2.2 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 $\mathbf{\Omega}_i$.

Consider the case where three measurements are taken by both methods A and B , \mathbf{y}_i is a 6×1 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.

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

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

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

Information on the fixed effects are contained in a three dimensional vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$. For computational purposes β_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 2×2 matrix, while \mathbf{R}_i is a $2n_i \times 2n_i$ matrix.

2.3 Decomposition of the response covariance matrix

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

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

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

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

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

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

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

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

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

The diagonal blocks are $\mathbf{\Sigma}$, as described previously. The 2×2 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 $\mathbf{\Lambda}$.

2.4 Sampling Scheme : Linked and Unlinked Replicates

Measurements taken in quick succession by the same observer using the same instrument on the same subject can be considered true replicates. Roy (2009b) 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.

One important feature of replicate observations is that they should be independent of each other. In essence, this is achieved by ensuring that the observer makes each measurement independent of knowledge of the previous value(s). This may be difficult to achieve in practice. (Check who said this)

2.5 Replicate measurements

Roy (2009b) accords with Bland and Altmans definition of a replicate, as being two or more measurements on the same individual under identical conditions. Roy allows the assumption that replicated measurements are equi-correlated. Roy allows unequal numbers of replicates.

Replicate measurements are linked over time. However the method can be easily extended to cover situations where they are not linked over time.

Chapter 3

Introduction to Roy's Procedure

Roy (2009b) proposes the use of LME models to perform a test on two methods of agreement to comparing the agreement between two methods of measurement, where replicate measurements on items (often individuals) by both methods are available, determining whether they can be used interchangeably. This approach uses a Kronecker product covariance structure with doubly multivariate setup to assess the agreement, and is designed such that the data may be unbalanced and with unequal numbers of replications for each subject (Roy, 2009b).

Three tests of hypothesis are provided, appropriate for evaluating the agreement between the two methods of measurement under this sampling scheme. These tests consider null hypotheses that assume: absence of inter-method bias; equality of between-subject variabilities of the two methods; equality of within-subject variabilities of the two methods. By inter-method bias we mean that a systematic difference exists between observations recorded by the two methods.

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

3.1 Introduction to Roy's methodology

Roy (2009b) uses an approach based on linear mixed effects (LME) models for the purpose of comparing the agreement between two methods of measurement, where replicate measurements on items (often individuals) by both methods are available. She provides three tests of hypothesis appropriate for evaluating the agreement between the two methods of measurement under this sampling scheme. These tests consider null hypotheses that assume: absence of inter-method bias; equality of between-subject variabilities of the two methods; equality of within-subject variabilities of the two methods. By inter-method bias we mean that a systematic difference exists between observations recorded by the two methods. Differences in between-subject variabilities of the two methods arise when one method is yielding average response levels for individuals 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.

For the purposes of comparing two methods of measurement, Roy (2009b) 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.

Roy (2009b) proposes the use of LME models to perform a test on two methods of agreement to determine whether they can be used interchangeably.

3.2 Model Set Up

Roy 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 (?).

For the purposes of comparing two methods of measurement, Roy (2009b) 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.

Roy (2009b) proposes the use of LME models to perform a test on two methods of agreement to comparing the agreement between two methods of measurement, where replicate measurements on items (often individuals) by both methods are available, determining whether they can be used interchangeably. This approach uses a Kronecker product covariance structure with doubly multivariate setup to assess the agreement, and is designed such that the data may be unbalanced and with unequal numbers of replications for each subject (Roy, 2009b).

The formulation contains a Kronecker product covariance structure in a doubly multivariate setup. By doubly multivariate set up, Roy means that the information on each patient or item is multivariate at two levels, the number of methods and number of replicated measurements. Further to Lam et al. (1999), it is assumed that the replicates are linked over time. However it is easy to modify to the unlinked case.

3.3 Criteria

Roy (2009b) sets out three criteria for two methods to be considered in agreement. Firstly that there be no significant bias. Second that there is no difference in the between-subject variabilities, and lastly that there is no significant difference in the

within-subject variabilities. Roy further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal.

3.4 Test for inter-method bias

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

A formal test for inter-method bias can be implemented by examining the fixed effects of the model. This is common to well known classical linear model methodologies. The null hypotheses, that both methods have the same mean, which is tested against the alternative hypothesis, that both methods have different means.

The inter-method bias and necessary t -value and p -value are presented in computer output. A decision on whether the first of Roy's criteria is fulfilled can be based on these values.

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

3.5 Variability Tests

Importantly Roy (2009b) further proposes a series of three tests on the variance components of an LME model, which allow decisions on the second and third of Barnhart's criteria. For these tests, four candidate LME models are constructed. The differences in the models are specifically in how the D and Λ matrices are constructed, using either an unstructured form or a compound symmetry form. To illustrate these differences, consider a generic matrix A ,

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

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

3.6 Roy's Candidate Models

Using Roy's method, four candidate models are constructed, each differing by constraints applied to the variance covariance matrices. In addition to computing the inter-method bias, three significance tests are carried out on the respective formulations to make a judgement on whether or not two methods are in agreement.

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

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

The likelihood function is a fundamental concept in statistical inference. It indicates how likely a particular population is to produce an observed sample. The set of values that maximize the likelihood function are considered to be optimal, and are used as the estimates of the parameters.

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

REML estimation reduces the bias in the variance component, and also handles high correlations more effectively, and is less sensitive to outliers than ML. The problem with REML for model building is that the "likelihoods" obtained for different fixed effects are not comparable. Hence it is not valid to compare models with different fixed effects using a likelihood ratio test or AIC when REML is used to estimate the model. Therefore models derived using ML must be used instead.

3.7 Likelihood Ratio Tests

The relationship between the respective models presented by Roy (2009b) is known as “nesting”. A model A to be nested in the reference model, model B, if Model A is a special case of Model B, or with some specific constraint applied.

A general method for comparing models with a nesting relationship is the likelihood ratio test (LRTs). LRTs are a family of tests used to compare the value of likelihood functions for two models, whose respective formulations define a hypothesis to be tested (i.e. the nested and reference model). The significance of the likelihood ratio test can be found by comparing the likelihood ratio to the χ^2 distribution, with the appropriate degrees of freedom.

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

3.7.1 Likelihood Ratio Tests

The relationship between the respective models presented by Roy (2009a) is known as “nesting”. A model A to be nested in the reference model, model B, if Model A is a special case of Model B, or with some specific constraint applied.

A general method for comparing models with a nesting relationship is the likelihood ratio test (LRTs). LRTs are a family of tests used to compare the value of likelihood functions for two models, whose respective formulations define a hypothesis to be tested (i.e. the nested and reference model). The significance of the likelihood ratio test can be found by comparing the likelihood ratio to the χ^2 distribution, with the appropriate degrees of freedom.

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

3.7.2 Likelihood Ratio Tests

The relationship between the respective models presented by Roy (2009a) is known as “nesting”. A model A to be nested in the reference model, model B, if Model A is a special case of Model B, or with some specific constraint applied.

A general method for comparing nested models fitted by ML is the ***likelihood ratio test*** (Cite: Lehmann 1986). Such a test can also be used for models fitted using REML, but only if both models have been fitted by REML, and if the fixed effects specification is the same for both models.

If k_i is the number of parameters to be estimated in model i , then the asymptotic, or “large sample”, distribution of the LRT statistic, under the null hypothesis that the restricted model is adequate, is a χ^2 distribution with $k_2 - k_1$ degrees of freedom (Pinheiro and Bates, 1994, pg.83).

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

A general method for comparing models with a nesting relationship is the likelihood ratio test (LRTs). LRTs are a family of tests used to compare the value of likelihood functions for two models, whose respective formulations define a hypothesis to be tested (i.e. the nested and reference model). The significance of the likelihood ratio test can be found by comparing the likelihood ratio to the χ^2 distribution, with the appropriate degrees of freedom.

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

3.8 Model Selection Using Likelihood Ratio Tests

An important step in the process of model selection is to determine, for a given pair of models, if there is a “nesting relationship” between the two.

We define Model A to be “nested” in Model B if Model A is a special case of Model B, i.e. Model B with a specific constraint applied.

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

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 LRT is the difference of the log-likelihood functions, multiplied by -2 . The probability distribution of the test statistic is approximated by the χ^2 distribution with $(\nu_1 - \nu_2)$ degrees of freedom, where ν_1 and ν_2 are the degrees of freedom of models 1 and 2 respectively.

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

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

and the maximum likelihood estimate is the solution to the score equation

$$S(\theta) = 0.$$

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

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

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

3.8.1 Likelihood Ratio Tests

The relationship between the respective models presented by Roy (2009a) is known as “nesting”. A model A to be nested in the reference model, model B, if Model A is a special case of Model B, or with some specific constraint applied.

A general method for comparing models with a nesting relationship is the likelihood ratio test (LRTs). LRTs are a family of tests used to compare the value of likelihood functions for two models, whose respective formulations define a hypothesis to be tested (i.e. the nested and reference model). The significance of the likelihood ratio test can be found by comparing the likelihood ratio to the χ^2 distribution, with the appropriate degrees of freedom.

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

3.9 Likelihood Ratio Tests

PB on LRTS for LMEs

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.

I think they are talking, especially, about situations in which the number of model parameter differences (differences between the complex model and the nested simpler model) is large relative to the number of observations.

This is not really a worry for this dataset, but I will come back to the substance of this view, and alternatives to the approach taken here.

3.9.1 Pinheiro Bates

A general method for comparing nested models fitted by ML is the *likelihood ratio test* (Cite: Lehmann 1986). Such a test can also be used for models fitted using REML, but only if both models have been fitted by REML, and if the fixed effects specification is the same for both models.

If k_i is the number of parameters to be estimated in model i , then the asymptotic, or “large sample”, distribution of the LRT statistic, under the null hypothesis that the restricted model is adequate, is a χ^2 distribution with $k_2 - k_1$ degrees of freedom (Pinheiro and Bates, 1994, pg.83).

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

3.9.2 Empirical p-values of LRT tests

For both REML and ML estimates, the nominal p -values for the LRT statistics under a χ^2 distribution with 2 degrees of freedom are much greater than empirical values.

A number of ways of dealing with this issues are discussed (Pinheiro and Bates, 1994, 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.(Pinheiro and Bates, 1994, pg.87).

3.9.3 Other material

A general method for comparing nested models fit by maximum likelihood is the *likelihood ratio test*. This test can be used for models fit by REML (restricted maximum 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).

- Example of a likelihood ratio test used to compare two models:

```
>anova(modelA, modelB)
```

- The output will contain a p-value, and this should be used in conjunction with the AIC scores to judge which model is preferred. Lower AIC scores are better.
- Generally, likelihood ratio tests should be used to evaluate the significance of terms on the random effects portion of two nested models, and should not be used to determine the significance of the fixed effects.
- A simple way to more reliably test for the significance of fixed effects in an LME model is to use conditional F-tests, as implemented with the simple “anova” function. Example:

```
>anova(modelA)
```

will give the most reliable test of the fixed effects included in model1.

3.9.4 Nested and Reference Models

Hypotheses can be formulated in the context of a pair of models that have a nesting relationship [CITE: West et al].

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

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

3.9.5 LRTs for covariance parameters

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

3.9.6 Likelihood Ratio Tests

The relationship between the respective models presented by Roy (2009a) is known as “nesting”. A model A to be nested in the reference model, model B, if Model A is a special case of Model B, or with some specific constraint applied.

A general method for comparing models with a nesting relationship is the likelihood ratio test (LRTs). LRTs are a family of tests used to compare the value of likelihood functions for two models, whose respective formulations define a hypothesis to be tested (i.e. the nested and reference model). The significance of the likelihood ratio test can be found by comparing the likelihood ratio to the χ^2 distribution, with the appropriate degrees of freedom.

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

- LMEs
- Likelihood and log likelihood functions
- Likelihood ratio test
- more on score functions etc
- MLEs
- Algorithms

The estimate for the fixed effects are referred to as the best linear unbiased estimates (BLUE). Henderson’s estimate for the random effects is known as the best linear unbiased predictor (BLUP).

3.9.7 Likelihood Ratio Tests

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

The power of the likelihood ratio test may depend on specific sample size and the specific number of replications, and [Roy 2009] proposes simulation studies to examine this further.

3.9.8 Relevance of Estimation Methods

Nested LME models, fitted by ML estimation, can be compared using the likelihood ratio test [Lehmann (1986)]. Models fitted using REML estimation can also be compared, but only if both were fitted using REML, and both have the same fixed effects specifications.

Likelihood ratio tests are generally used to test the significance of terms in the random effects structure. Information Criteria Additionally nested models may be compared by using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

When comparing the respective scores for nested models, the model with the smaller score is considered to be the preferable model. ML / REML [Morrell 1998] The variance components in the LME model may be estimated by ML or REML. Maximum Likelihood estimates do not take into account the estimation of fixed effects and so are biased downwards. REML estimates account for the presence of these nuisance parameters by maximising the linearly independent error contrasts to obtain more unbiased estimates. Treatment of items as fixed effects [Pinheiro Bates 2000] addresses the issue of treating items as fixed effects. Such a specification is useful only for the specific sample of items, rather than the population of items, where the interest would naturally lie.

[Pinheiro Bates 2000] advises the specification of random effects to correspond to items; treating the item effects as random deviations from the population mean.

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

The log likelihood $l(\theta)$

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

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

3.9.9 Other material

A general method for comparing nested models fit by maximum likelihood is the *likelihood ratio test*. This test can be used for models fit by REML (restricted maximum 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).

- Example of a likelihood ratio test used to compare two models:

```
>anova(modelA, modelB)
```

- The output will contain a p-value, and this should be used in conjunction with the AIC scores to judge which model is preferred. Lower AIC scores are better.
- Generally, likelihood ratio tests should be used to evaluate the significance of terms on the random effects portion of two nested models, and should not be used to determine the significance of the fixed effects.
- A simple way to more reliably test for the significance of fixed effects in an LME model is to use conditional F-tests, as implemented with the simple “anova” function. Example:

```
>anova(modelA)
```

will give the most reliable test of the fixed effects included in model1.

3.9.10 Nested and Reference Models

Hypotheses can be formulated in the context of a pair of models that have a nesting relationship [CITE: West et al].

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

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

3.9.11 Test for inter-method bias

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

3.10 Likelihood Ratio Tests

The first model acts as an alternative hypothesis to be compared against each of three other models, acting as null hypothesis models, successively. The models are compared using the likelihood ratio test. Likelihood ratio tests are a class of tests based on the comparison of the values of the likelihood functions of two candidate models. LRTs can be used to test hypotheses about covariance parameters or fixed effects parameters in the context of LMEs. The test statistic for the likelihood ratio test is the difference of the log-likelihood functions, multiplied by -2 . The probability distribution of the test statistic is approximated by the χ^2 distribution with $(\nu_1 - \nu_2)$ degrees of freedom, where ν_1 and ν_2 are the degrees of freedom of models 1 and 2 respectively. Each of these three test shall be examined in more detail shortly.

3.11 Roy's LME methodology for assessing agreement

Three tests of hypothesis are provided, appropriate for evaluating the agreement between the two methods of measurement under this sampling scheme. These tests consider null hypotheses that assume: absence of inter-method bias; equality of between-subject variabilities of the two methods; equality of within-subject variabilities of the two methods. By inter-method bias we mean that a systematic difference exists between observations recorded by the two methods. Differences in between-subject variabilities of the two methods arise when one method is yielding average response levels for individuals 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.

Chapter 4

Model Specification

4.1 Model Specification for Roy's Hypotheses Tests

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

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

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ 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 $\boldsymbol{\epsilon}_i$ is a $2n_i \times 1$ vector of residual terms. Random effects and residuals are assumed to be independent of each other.

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

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

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

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

It is assumed that $\mathbf{b}_i \sim N(0, \mathbf{G})$, $\boldsymbol{\epsilon}_i$ 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{G} and \mathbf{R}_i will be discussed in due course.

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

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

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

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

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

where σ_1^2 and σ_2^2 are the within-subject variances of the respective methods, and σ_{12} is the within-item covariance between the two methods. The within-item variance covariance matrix $\boldsymbol{\Sigma}$ is assumed to be the same for all replications. Computational analysis of linear mixed effects models allow for the explicit analysis of both \mathbf{G} and \mathbf{R}_i .

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

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

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

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

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

4.2 G Component

\mathbf{G} is the variance covariance matrix for the random effects \mathbf{b} . i.e. between-item sources of variation. It is important to note that no special assumptions about the structure of \mathbf{G} are made. An example of such an assumption would be that \mathbf{G} is the product of a scalar value and the identity matrix.

It is assumed that $\mathbf{b}_i \sim N(0, \mathbf{G})$, $\boldsymbol{\epsilon}_i$ 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{G} and \mathbf{R}_i will be discussed in due course.

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

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

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

4.3 R Component

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

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

where σ_1^2 and σ_2^2 are the within-subject variances of the respective methods, and σ_{12} is the within-item covariance between the two methods. The within-item variance covariance matrix $\boldsymbol{\Sigma}$ is assumed to be the same for all replications. Again it is important

to note that no special assumptions are made about the structure of the matrix. Computational analysis of linear mixed effects models allow for the explicit analysis of both \mathbf{G} and \mathbf{R}_i .

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

Computational analysis of linear mixed effects models allow for the explicit analysis of both \mathbf{G} and \mathbf{R}_i . The above terms can be used to express the variance covariance matrix $\mathbf{\Omega}_i$ for the responses on item i ,

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

The partial within-item variance covariance matrix of two methods at any replicate is denoted $\mathbf{\Sigma}$, where σ_1^2 and σ_2^2 are the within-subject variances of both methods, and σ_{12} is the within-item covariance between the two methods. The within-item variance covariance matrix $\mathbf{\Sigma}$ is assumed to be the same for all replications.

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

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

The computation of the limits of agreement require that the variance of the difference of measurements. This variance is easily computable from the estimate of the Block - $\mathbf{\Omega}_i$ matrix. Lack of agreement can arise if there is a disagreement in overall

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

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

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

where σ_1^2 and σ_2^2 are the within-subject variances of the respective methods, and σ_{12} is the within-item covariance between the two methods. The within-item variance covariance matrix $\boldsymbol{\Sigma}$ is assumed to be the same for all replications. Computational analysis of linear mixed effects models allow for the explicit analysis of both \mathbf{G} and \mathbf{R}_i .

The distribution of the random effects is described as $\mathbf{b}_i \sim N(0, \mathbf{G})$. Similarly random errors are distributed as $\boldsymbol{\epsilon}_i \sim N(0, \mathbf{R}_i)$. The random effects and residuals are assumed to be independent. Both covariance matrices can be written as follows;

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

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

4.4 Hamlett

\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{G} and \mathbf{R}_i .

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

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

Hamlett et al. (2004) shows that \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 σ_1^2 and σ_2^2 are the within-subject variances of the respective methods, and σ_{12} is the within-item covariance between the two methods. It is assumed that the within-item variance?covariance matrix $\mathbf{\Sigma}$ is the same for all replications. Again it is important to note that no special assumptions are made about the structure of the matrix.

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \quad (4.1)$$

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

4.5 For Expository Purposes

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

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

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

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

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

4.6 Kroneckor

The between-item variance covariance matrix \mathbf{G} is constructed as follows:

Both covariance matrices can be written as follows;

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

and

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

Hamlett et al. (2004) 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 σ_1^2 and σ_2^2 are the within-subject variances of the respective methods, and

σ_{12} is the within-item covariance between the two methods. It is assumed that the within-item variance?covariance matrix Σ is the same for all replications. Again it is important to note that no special assumptions are made about the structure of the matrix.

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \quad (4.2)$$

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

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

4.7 Overall Variability

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

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

The variance of case-wise difference in measurements can be determined from Block- Ω_i . Hence limits of agreement can be computed.

The computation of the limits of agreement require that the variance of the difference of measurements. This variance is easily computable from the estimate of the Block - Ω_i matrix. Lack of agreement can arise if there is a disagreement in overall

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

Chapter 5

Roy's hypothesis tests

5.1 Hypothesis Testing

Variability tests proposed by Roy (2009b) affords the opportunity to expand upon Carstensen's approach. Roy (2009b) considers four independent hypothesis tests. The first test allows of the comparison the begin-subject variability of two methods. Similarly, the second test assesses the within-subject variability of two methods. A third test is a test that compares the overall variability of the two methods.

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

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

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

- A formal test for the equality of overall variances.

These tests are complemented by the ability to consider the inter-method bias and the overall correlation coefficient. Two methods can be considered to be in agreement if criteria based upon these methodologies are met. Additionally Roy makes reference to the overall correlation coefficient of the two methods, which is determinable from variance estimates.

5.2 Roy's hypothesis tests

The presence of an inter-method bias is the source of disagreement between two methods of measurement that is most easily identified. As the first in a series of hypothesis tests, Roy (2009a) presents a formal test for inter-method bias. With the null and alternative hypothesis denoted H_1 and K_1 respectively, this test is formulated as

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

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

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,

$$H_2 : g_1^2 = g_2^2$$

$$K_2 : g_1^2 \neq g_2^2$$

and a formal test for the equality of within-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 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.

Conversely, the tests of variability required detailed explanation. Each test is performed by fitting two candidate models, according with the null and alternative hypothesis respectively. The distinction between the models arise in the specification in one, or both, of the variance-covariance matrices.

5.3 Roy's variability tests

The tests are implemented by fitting a specific LME model, and three variations thereof, to the data. These three variant models introduce equality constraints that act null hypothesis cases.

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

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

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

$$H_0 : g_1^2 = g_2^2$$

$$H_1 : g_1^2 \neq g_2^2$$

a formal test for the equality of within-item variances,

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

and finally, a formal test for the equality of overall variances.

$$H_0 : \omega_1^2 = \omega_2^2$$

$$H_1 : \omega_1^2 \neq \omega_2^2$$

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.

5.4 Variability test 1

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

$$H_0 : d_A = d_B$$

$$H_A : d_A \neq d_B$$

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

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

5.5 Variability test 2

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

$$H_0 : \lambda_A = \lambda_B$$

$$H_A : \lambda_A = \lambda_B$$

This model is performed in the same manner as the first test, only reversing the roles of $\hat{\mathbf{D}}$ and $\hat{\mathbf{\Lambda}}$. The null model is constructed a symmetric form for $\hat{\mathbf{\Lambda}}$ while the alternative model uses a compound symmetry form. This time $\hat{\mathbf{D}}$ 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, Roy (2009a) includes the coefficients of repeatability for both methods.

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

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

5.6 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.

$$H_0 : \sigma_A = \sigma_B$$

$$H_A : \sigma_A \neq \sigma_B$$

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

5.7 Omnibus testing

$$\begin{pmatrix} \omega_e^2 & \omega^{en} \\ \omega_{en} & \omega_n^2 \end{pmatrix} = \begin{pmatrix} \psi_e^2 & \psi^{en} \\ \psi_{en} & \psi_n^2 \end{pmatrix} + \begin{pmatrix} \sigma_e^2 & \sigma^{en} \\ \sigma_{en} & \sigma_n^2 \end{pmatrix} \quad (5.1)$$

$$\begin{pmatrix} \omega_2^1 & 0 \\ 0 & \omega_2^2 \end{pmatrix} = \begin{pmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

The computation of the limits of agreement require that the variance of the difference of measurements. This variance is easily computable from the estimate of the Block - Ω_i matrix. Lack of agreement can arise if there is a disagreement in overall variabilities. This may be due to the disagreement in either between-item variabilities or within-item variabilities, or both. Roy (2009b) allows for a formal test of each.

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \quad (5.2)$$

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

5.8 Implementation

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

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

5.9 Formal testing for covariances

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

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

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

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.

5.9.1 Variance Covariance Matrices

Under Roy's model, random effects are defined using a bivariate normal distribution. Consequently, the variance-covariance structures can be described using 2×2 matrices. A discussion of the various structures a variance-covariance matrix can be specified

under is required before progressing. The following structures are relevant: the identity structure, the compound symmetric structure and the symmetric structure.

The identity structure is simply an abstraction of the identity matrix. The compound symmetric structure and symmetric structure can be described with reference to the following matrix (here in the context of the overall covariance Block- $\mathbf{\Omega}_i$, but equally applicable to the component variabilities \mathbf{G} 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 ψ^2) for all 3 of the time points and a single covariance (represented by ψ_{ij}) for each of the pairs of trials.

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

Unstructured

Assumes that each variance and covariance is unique. Each trial has its own variance (e.g. s_{12} is the variance of trial 1) and each pair of trials has its own covariance (e.g. s_{21} is the covariance of trial 1 and trial2). This structure is illustrated by the half matrix below.

Autoregressive

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

5.10 VC structures

There is three alternative structures for Ψ , the diagonal form, the identity form and the general form.

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

Ψ is the variance-covariance matrix of the random effects , with 2×2 dimensions.

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

There is three alternative structures for Ψ , the diagonal form, the identity form and the general form.

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

Ψ is the variance-covariance matrix of the random effects , with 2×2 dimensions.

$$\Psi = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix} \tag{5.5}$$

Chapter 6

LOAS

6.1 Calculation of limits of agreement

However, the original BlandAltman method was developed for two sets of measurements done on one occasion (i.e. independent data), and so this approach is not suitable for replicate measures data. However, as a naive analysis, it may be used to explore the data because of the simplicity of the method.

The limits of agreement (Bland and Altman, 1986) are ubiquitous in method comparison studies.

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.

Computing limits of agreement features prominently in many method comparison studies, further to Bland and Altman (1986, 1999).

Bland and Altman (1999) addresses the issue of computing LoAs in the presence of replicate measurements, suggesting several computationally simple approaches. When repeated measures data are available, it is desirable to use all the data to compare the

two methods.

However, the original Bland-Altman method was developed for two sets of measurements done on one occasion (i.e. independent data), and so this approach is not suitable for replicate measures data. However, as a naive analysis, it may be used to explore the data because of the simplicity of the method. Carstensen et al. (2008) computes the limits of agreement to the case with replicate measurements by using LME models.

Further to Bland and Altman (1986), the computation of the limits of agreement follows from the intermethod bias, and the variance of the difference of measurements. The computation of the inter-method bias is a straightforward subtraction calculation. The variance of differences is easily computable from the variance estimates in the Block - Ω_i matrix, i.e.

$$\text{Var}(y_1 - y_2) = \sqrt{\omega_1^2 + \omega_2^2 - 2\omega_{12}}.$$

Carstensen et al. (2008) demonstrate statistical flaws with two approaches proposed by Bland and Altman (1999) for the purpose of calculating the variance of the inter-method bias when replicate measurements are available. Instead, Carstensen et al. (2008) use a fitted mixed effects model to obtain appropriate estimates for the variance of the inter-method bias. As their interest mainly lies in extending the Bland-Altman methodology, other formal tests are not considered.

Carstensen et al. (2008) also presents a methodology to compute the limits of agreement based on LME models. In many cases the limits of agreement derived from this method accord with those to Roy's model. However, in other cases dissimilarities emerge. An explanation for this differences can be found by considering how the respective models account for covariance in the observations. Specifying the relevant terms using a bivariate normal distribution, Roy's model allows for both between-method and within-method covariance. Carstensen et al. (2008) formulate a model whereby random effects have univariate normal distribution, and no allowance is made for correlation between observations.

A consequence of this is that the between-method and within-method covariance are zero. In cases where there is negligible covariance between methods, both sets of limits of agreement are very similar to each other. In cases where there is a substantial level of covariance present between the two methods, the limits of agreement computed using models will differ.

6.2 00-Limits of Agreement in LME models

Carstensen et al. (2008) computes the limits of agreement to the case with repeated measurements by using LME models.

Roy (2009b) formulates a very powerful method of assessing whether two methods of measurement, with replicate measurements, also using LME models. Roy's approach is based on the construction of variance-covariance matrices.

Importantly, Roy's approach does not address the issue of limits of agreement (though another related analysis, the coefficient of repeatability, is mentioned).

This paper seeks to use Roy's approach to estimate the limits of agreement. These estimates will be compared to estimates computed under Carstensen's formulation.

In computing limits of agreement, it is first necessary to have an estimate for the standard deviations of the differences. When the agreement of two methods is analyzed using LME models, a clear method of how to compute the standard deviation is required. As the estimate for inter-method bias and the quantile would be the same for both methodologies, the focus hereon is solely on the variance of differences.

6.3 Carstensen et al

Carstensen et al. (2008) also use a LME model for the purpose of comparing two methods of measurement where replicate measurements are available on each item. Their interest lies in generalizing the popular limits-of-agreement (LOA) methodology advocated by Bland and Altman (1986) to take proper cognizance of the replicate mea-

surements. Carstensen et al. (2008) demonstrate statistical flaws with two approaches proposed by Bland and Altman (1999) for the purpose of calculating the variance of the inter-method bias when replicate measurements are available. Instead, they recommend a fitted mixed effects model to obtain appropriate estimates for the variance of the inter-method bias. As their interest mainly lies in extending the Bland-Altman methodology, other formal tests are not considered.

6.4 Computing LoAs from LME models

One important feature of replicate observations is that they should be independent of each other. In essence, this is achieved by ensuring that the observer makes each measurement independent of knowledge of the previous value(s). This may be difficult to achieve in practice.

6.5 Carstensen’s Limits of Agreement

Carstensen et al. (2008) presents a methodology to compute the limits of agreement based on LME models. The method of computation is the same as Roy’s model, but with the covariance estimates set to zero.

Importantly, Carstensen’s underlying model differs from Roy’s model in some key respects, and therefore a prior discussion of Carstensen’s model is required.

Carstensen et al. (2008) presents a methodology to compute the limits of agreement based on LME models. Importantly, Carstensen’s underlying model differs from Roy’s model in some key respects, and therefore a prior discussion of Carstensen’s model is required. The method of computation is the same as Roy’s model, but with the covariance estimates set to zero.

Carstensen et al. (2008) uses LME models to determine the limits of agreement. In computing limits of agreement, it is first necessary to have an estimate for the standard deviations of the differences. When the agreement of two methods is analyzed using

LME models, a clear method of how to compute the standard deviation is required. As the estimate for inter-method bias and the quantile would be the same for both methodologies, the focus is solely on the standard deviation.

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

Importantly, estimates required to calculate the limits of agreement are not extractable, and therefore the calculation must be done by hand.

Carstensen presents a model where the variation between items for method m is captured by σ_m and the within item variation by τ_m . Further to his model, Carstensen computes the limits of agreement as

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

Further to Bland and Altman (1986), the computation of the limits of agreement follows from the intermethod bias, and the variance of the difference of measurements. The computation thereof require that the variance of the difference of measurements. This variance is easily computable from the variance estimates in the Block - $\mathbf{\Omega}_i$ matrix, i.e.

$$\text{Var}(y_1 - y_2) = \sqrt{\omega_1^2 + \omega_2^2 - 2\omega_{12}}.$$

6.6 Carstensen's Model

Using Carstensen's notation, a measurement y_{mi} by method m on individual i the measurement y_{mir} is the r th replicate measurement on the i th item by the m th method, where $m = 1, 2$, $i = 1, \dots, N$, and $r = 1, \dots, n_i$ is formulated as follows;

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + \epsilon_{mir}, \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \quad (6.1)$$

Of particular importance is terms of the model, a true value for item i (μ_i). The fixed effect of Roy's model comprise of an intercept term and fixed effect terms for both methods, with no reference to the true value of any individual item. A distinction can be made between the two models: Roy's model is a standard LME model, whereas Carstensen's model is a more complex additive model.

The classical model is based on measurements y_{mi} by method $m = 1, 2$ on item $i = 1, 2, \dots$

$$y_{mi} + \alpha_m + \mu_i + e_{mi}$$

$$e_{mi} \sim N(0, \sigma_m^2)$$

Here the terms α_m and μ_i represent the fixed effect for method m and a true value for item i respectively. The random effect terms comprise an interaction term c_{mi} and the residuals ϵ_{mir} . The c_{mi} term represent random effect parameters corresponding to the two methods, having $E(c_{mi}) = 0$ with $\text{Var}(c_{mi}) = \tau_m^2$. Carstensen specifies the variance of the interaction terms as being univariate normally distributed. As such, $\text{Cov}(c_{mi}, c_{m'i}) = 0$. All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method.

Even though the separate variances can not be identified, their sum can be estimated by the empirical variance of the differences.

Like wise the separate α can not be estimated, only their difference can be estimated as \bar{D}

The presence of the true value term μ_i gives rise to an important difference between Carstensen's and Roy's models. The fixed effect of Roy's model comprise of an intercept term and fixed effect terms for both methods, with no reference to the true value of any individual item. In other words, Roy considers the group of items being measured as a sample taken from a population. Therefore a distinction can be made between the

two models: Roy's model is a standard LME model, whereas Carstensen's model is a more complex additive model.

With regards to specifying the variance terms, Carstensen remarks that using his approach is common, remarking that *The only slightly non-standard (meaning "not often used") feature is the differing residual variances between methods* (Carstensen, 2010).

Carstensen et al. (2008) makes some interesting remarks in this regard.

The only slightly non-standard (meaning "not often used") feature is the differing residual variances between methods.

Further to his model, Carstensen computes the limits of agreement as

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

As the difference between methods is of interest, the item term can be disregarded.

We assume that the variance of the measurements is different for both methods, but it does not mean that the separate variances can be estimated with the data available.

6.7 BXC2008 presents - Carstensen's Limits of agreement

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

Importantly, estimates required to calculate the limits of agreement are not extractable, and therefore the calculation must be done by hand.

Carstensen et al. (2008) use a LME model for the purpose of comparing two methods of measurement where replicate measurements are available on each item. Their interest lies in generalizing the popular limits-of-agreement (LOA) methodology advocated by Bland and Altman (1986) to take proper cognizance of the replicate measurements. Carstensen et al. (2008) demonstrate statistical flaws with two approaches proposed by Bland and Altman (1999) for the purpose of calculating the variance of the inter-method bias when replicate measurements are available, instead proposing a fitted mixed effects model to obtain appropriate estimates for the variance of the inter-method bias. As their interest lies specifically in extending the Bland-Altman methodology, other formal tests are not considered.

Carstensen et al. (2008) also presents a methodology to compute the limits of agreement based on LME models. The method of computation is similar Roy's model, but

for absence of the covariance estimates. In cases where there is negligible covariance between methods, the limits of agreement computed using Roy's model accord with those computed using model described by (1.4). In cases where some degree of covariance is present between the two methods, the limits of agreement computed using models will differ. In the presented example, it is shown that Roy's LOAs are lower than those of (1.4), when covariance between methods is present.

Carstensen et al. (2008) presents a methodology to compute the limits of agreement based on LME models. Importantly, Carstensen's underlying model differs from Roy's model in some key respects, and therefore a prior discussion of Carstensen's model is required.

6.8 Limits of Agreement in LME models

Carstensen et al. (2008) 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 λ_m^2 . Carstensen et al. (2008) 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

$$\text{var}(y_A - y_B) = 2d^2 + \lambda_A^2 + \lambda_B^2. \quad (6.2)$$

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

Roy (2009b) 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

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_A^2 + \lambda_B^2 - 2(d_{AB} + \lambda_{AB}) \quad (6.3)$$

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

$$\bar{y}_A - \bar{y}_B \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \quad (6.4)$$

6.9 Limits of Agreement in LME models

Carstensen et al. (2008) uses LME models to determine the limits of agreement. The limits of agreement (Bland and Altman, 1986) are ubiquitous in method comparison studies.

6.10 BXC2008 uses - Limits of Agreement in LME models

Carstensen et al. (2008) uses an approach based on linear mixed effects (LME) models for the purpose of computing the limits of agreement for two methods of measurement, where replicate measurements are taken on items. As the emphasis of this methodology lies on the inter-method bias and the limits of agreement, the two key elements of the Bland-Altman methodology, other formal tests are not described.

Between-subject variation for method m is given by d_m^2 and within-subject variation is given by λ_m^2 . Carstensen et al. (2008) 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

$$\text{var}(y_A - y_B) = 2d^2 + \lambda_A^2 + \lambda_B^2. \quad (6.5)$$

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

6.11 Carstensen's LOAs

Carstensen presents a model where the variation between items for method m is captured by σ_m and the within item variation by τ_m .

Further to his model, Carstensen computes the limits of agreement as

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

The respective estimates computed by both methods are tabulated as follows. Evidently there is close correspondence between both sets of estimates.

6.12 Interaction Terms in Model

Carstensen et al. (2008) formulates an LME model, both in the absence and the presence of an interaction term. ? uses both to demonstrate the importance of using an interaction term. Failure to take the replication structure into account results in overestimation of the limits of agreement. For the Carstensen estimates below, an interaction term was included when computed.

Carstensen presents a model where the variation between items for method m is captured by σ_m and the within item variation by τ_m .

Further to his model, Carstensen computes the limits of agreement as

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

6.13 Computation of limits of agreement

The computation thereof require that the variance of the difference of measurements. This variance is easily computable from the variance estimates in the Block - Ω_i matrix, i.e.

$$\text{Var}(y_1 - y_2) = \sqrt{\omega_1^2 + \omega_2^2 - 2\omega_{12}}.$$

Carstensen et al. (2008) also presents a methodology to compute the limits of agreement based on LME models. The method of computation is similar Roy's model, but for absence of the covariance estimates. In cases where there is negligible covariance between methods, the limits of agreement computed using Roy's model accord with those computed using model described by (1.4). In cases where some degree of covariance is present between the two methods, the limits of agreement computed using models will differ. In the presented example, it is shown that Roy's LOAs are lower than those of (1.4), when covariance between methods is present.

6.14 BXC - Model Terms

- Let y_{mir} be the response of method m on the i th subject at the r -th replicate.
- Let \mathbf{y}_{ir} be the 2×1 vector of measurements corresponding to the i -th subject at the r -th replicate.
- Let \mathbf{y}_i be the $R_i \times 1$ vector of measurements corresponding to the i -th subject, where R_i is number of replicate measurements taken on item i .

- Let α_{mi} be the fixed effect parameter for method for subject i .
- Formally Roy uses a separate fixed effect parameter to describe the true value μ_i , but later combines it with the other fixed effects when implementing the model.
- Let u_{1i} and u_{2i} be the random effects corresponding to methods for item i .
- ϵ_i is a n_i -dimensional vector comprised of residual components. For the blood pressure data $n_i = 85$.
- β is the solutions of the means of the two methods. In the LME output, the bias and corresponding t-value and p-values are presented. This is relevant to Roy's first test.

6.15 Computation of limits of agreement under Roy's model

The limits of agreement computed by Roy's method are derived from the variance covariance matrix for overall variability. This matrix is the sum of the between subject VC matrix and the within-subject VC matrix. The computation thereof require that the variance of the difference of measurements. This variance is easily computable from the variance estimates in the Block - Ω_i matrix, i.e.

$$\text{Var}(y_1 - y_2) = \sqrt{\omega_1^2 + \omega_2^2 - 2\omega_{12}}.$$

The standard deviation of the differences of methods x and y is computed using values from the overall VC matrix.

$$\text{var}(x - y) = \text{var}(x) + \text{var}(y) - 2\text{cov}(x, y)$$

The respective estimates computed by both methods are tabulated as follows. Evidently there is close correspondence between both sets of estimates.

6.16 LOAs with Roy

Roy (2009b) 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

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_A^2 + \lambda_B^2 - 2(d_{AB} + \lambda_{AB}) \quad (6.6)$$

The limits of agreement computed by Roy's method are derived from the variance covariance matrix for overall variability. This matrix is the sum of the between subject VC matrix and the within-subject VC matrix.

The standard deviation of the differences of methods x and y is computed using values from the overall VC matrix.

$$\text{var}(x - y) = \text{var}(x) + \text{var}(y) - 2\text{cov}(x, y)$$

The respective estimates computed by both methods are tabulated as follows. Evidently there is close correspondence between both sets of estimates.

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

$$\bar{y}_A - \bar{y}_B \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \quad (6.7)$$

The computation thereof require that the variance of the difference of measurements. This variance is easily computable from the variance estimates in the Block - $\mathbf{\Omega}_i$ matrix, i.e.

$$\text{Var}(y_1 - y_2) = \sqrt{\omega_1^2 + \omega_2^2 - 2\omega_{12}}.$$

6.17 Differences Between Models

Carstensen et al. (2008) also presents a methodology to compute the limits of agreement based on LME models. In many cases the limits of agreement derived from this method

accord with those to Roy's model. However, in other cases dissimilarities emerge. An explanation for this differences can be found by considering how the respective models account for covariance in the observations. Specifying the relevant terms using a bivariate normal distribution, Roy's model allows for both between-method and within-method covariance. Carstensen et al. (2008) formulate a model whereby random effects have univariate normal distribution, and no allowance is made for correlation between observations.

A consequence of this is that the between-method and within-method covariance are zero. In cases where there is negligible covariance between methods, both sets of limits of agreement are very similar to each other. In cases where there is a substantial level of covariance present between the two methods, the limits of agreement computed using models will differ.

There is a substantial difference in the number of fixed parameters used by the respective models. For the model in (2.3) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items N . In contrast, the model described by (1.4) requires $N + 2$ fixed effects for N items. The inclusion of fixed effects to account for the 'true value' of each item greatly increases the level of model complexity.

When only two methods are compared, Carstensen et al. (2008) notes that separate estimates of τ_m^2 can not be obtained due to the model over-specification. To overcome this, the assumption of equality, i.e. $\tau_1^2 = \tau_2^2$, is required.

6.18 Differences

Roy (2009b) 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

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_B^2 + \lambda_A^2 - 2(d_{AB} + \lambda_{AB}) \quad (6.8)$$

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

$$\bar{y}_A - \bar{y}_B \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \quad (6.9)$$

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

Importantly, estimates required to calculate the limits of agreement are not extractable, and therefore the calculation must be done by hand. Carstensen presents a model where the variation between items for method m is captured by σ_m and the within item variation by τ_m .

In contrast to Roy's model, Carstensen's model requires that commonly used assumptions be applied, specifically that the off-diagonal elements of the between-item and within-item variability matrices are zero. By extension the overall variability off-diagonal elements are also zero. Also, implementation requires that the between-item variances are estimated as the same value: $g_1^2 = g_2^2 = g^2$. As a consequence, Carstensen's method does not allow for a formal test of the between-item variability.

$$\begin{pmatrix} \omega_2^1 & 0 \\ 0 & \omega_2^2 \end{pmatrix} = \begin{pmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

6.19 Relevance of Roy's Methodology

The relevance of Roy's methodology is that estimates for the between-item variances for both methods \hat{d}_m^2 are computed. Also the VC matrices are constructed with covariance terms and, so the difference variance must be formulated accordingly.

$$\hat{\alpha}_1 - \hat{\alpha}_2 \pm \sqrt{\hat{d}_1^2 + \hat{d}_2^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{d}_{12} - 2\hat{\sigma}_1\hat{\sigma}_2}$$

Roy (2009b) considers the problem of assessing the agreement between two methods with replicate observations in a doubly multivariate set-up using linear mixed effects models.

Roy (2009b) uses examples from Bland and Altman (1986) to be able to compare both types of analysis.

Roy (2009b) proposes a LME based approach with Kronecker product covariance structure with doubly multivariate setup to assess the agreement between two methods. This method is designed such that the data may be unbalanced and with unequal numbers of replications for each subject.

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

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \tag{6.10}$$

6.20 Difference Variance further to Carstensen

Carstensen et al. (2008) states a model where the variation between items for method m is captured by τ_m (our notation d_m^2) and the within-item variation by σ_m .

The formulation of this model is general and refers to comparison of any number of methods however, if only two methods are compared, separate values of τ_1^2 and τ_2^2

cannot be estimated, only their average value τ , so in the case of only two methods we are forced to assume that $\tau_1 = \tau_2 = \tau$ (Carstensen et al., 2008).

Another important point is that there is no covariance terms, so further to Carstensen et al. (2008) the variance covariance matrices for between-item and within-item variability are respectively.

$$\mathbf{D} = \begin{pmatrix} d_2^1 & 0 \\ 0 & d_2^2 \end{pmatrix}$$

and $\mathbf{\Sigma}$ is constructed as follows:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_2^1 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

Under this model the limits of agreement should be computed based on the standard deviation of the difference between a pair of measurements by the two methods on a new individual, j , say:

$$\text{var}(y_{1j} - y_{2j}) = 2d^2 + \sigma_1^2 + \sigma_2^2$$

Further to his model, Carstensen computes the limits of agreement as

$$\hat{\alpha}_1 - \hat{\alpha}_2 \pm \sqrt{2\hat{d}^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2}$$

6.21 Assumptions on Variability

Aside from the fixed effects, another important difference is that Carstensen's model requires that particular assumptions be applied, specifically that the off-diagonal elements of the between-item and within-item variability matrices are zero. By extension the overall variability off diagonal elements are also zero.

Also, implementation requires that the between-item variances are estimated as the same value: $g_1^2 = g_2^2 = g^2$. Necessarily Carstensen's method does not allow for a formal test of the between-item variability.

$$\begin{pmatrix} \omega_2^1 & 0 \\ 0 & \omega_2^2 \end{pmatrix} = \begin{pmatrix} g^2 & 0 \\ 0 & g^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

In cases where the off-diagonal terms in the overall variability matrix are close to zero, the limits of agreement due to ? are very similar to the limits of agreement that follow from the general model.

Chapter 7

Other Material

7.1 Extension of Roy's methodology

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

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

Should an implementation be feasible, further difficulty arises when interpreting the results. The fundamental question is whether two methods have close agreement so as to be interchangeable. When three methods are present in the model, the null

hypothesis is that all three methods have the same variability relevant to the respective tests. The outcome of the analysis will either be that all three are interchangeable or that all three are not interchangeable.

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

Roy's methodology is not suitable for the case of single measurements because it follows from the decomposition for the covariance matrix of the response vector y_i , as presented in Hamlett et al. (2004). The decomposition depends on the estimation of correlation terms, which would be absent in the single measurement case. Indeed there can be no within-subject variability if there are no repeated terms for it to describe. There would only be the covariance matrix of the measurements by both methods, which doesn't require the use of LME models. To conclude, simpler existing methodologies, such as Deming regression, would be the correct approach where there only one measurements by each method.

7.1.1 Correlation

Bivariate correlation coefficients have been shown to be of limited use in method comparison studies (Bland and Altman, 1986). However, recently correlation analysis has been developed to cope with repeated measurements, enhancing their potential usefulness. Roy incorporates the use of correlation into his methodology.

In addition to the variability tests, Roy advises that it is preferable that a correlation of greater than 0.82 exist for two methods to be considered interchangeable. However if two methods fulfil all the other conditions for agreement, failure to comply with this one can be overlooked. Indeed Roy demonstrates that placing undue importance to it can lead to incorrect conclusions. Roy (2009b) 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.

7.1.2 Correlation terms

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

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

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

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

7.2 Hamlett and Lam

The methodology proposed by ? is largely based on Hamlett et al. (2004), which in turn follows on from Lam et al. (1999).

Hamlett re-analyses the data of Lam et al. (1999) to generalize their model to cover other settings not covered by the Lam method.

In many cases, repeated observation are collected from each subject in sequence and/or longitudinally.

$$y_i = \alpha + \mu_i + \epsilon$$

7.3 LaiShiao

Lai and Shiao (2005) advocates the use of LME models to study method comparison problems. The authors analyse a data set typical of method comparison studies using SAS software, with particular use of the ‘*Proc Mixed*’ package. The stated goal of this study is to determine which factor from a specified group of factors is the key contributor to the difference in the two methods.

The study relates to oxygen saturation, the most investigated variable in clinical nursing studies (Lai and Shiao, 2005). The two method compared are functional saturation (SO₂, percent functional oxy-hemoglobin) and fractional saturation (HbO₂, percent fractional oxy-hemoglobin), which is considered to be the ‘gold standard’ method of measurement.

Lai and Shiao (2005) establishes an LME model for analysing the differences D_{ijtl} , where D_{ijtl} is the differences of the measurements (i.e. $SO_{2_{ijtl}} - HbO_{2_{ijtl}}$) for the i th donor at the j th level of foetal haemoglobin percent (Fhbperct) and the t th repeated measurement by the l th practitioner of the experiment.

(Carstensen (2004) also advocates the use of LME models in comparing methods, but with a different emphasis.)

7.4 LaiShiao

Lai and Shiao (2005) advocates the use of LME models to study method comparison problems. The authors analyse a data set typical of method comparison studies using SAS software, with particular use of the ‘*Proc Mixed*’ package. The stated goal of this study is to determine which factor from a specified group of factors is the key contributor to the difference in the two methods.

The study relates to oxygen saturation, the most investigated variable in clinical nursing studies (Lai and Shiao, 2005). The two method compared are functional saturation (SO₂, percent functional oxy-hemoglobin) and fractional saturation (HbO₂, percent fractional oxy-hemoglobin), which is considered to be the ‘gold standard’ method of measurement.

Lai and Shiao (2005) establishes an LME model for analysing the differences D_{ijtl} , where D_{ijtl} is the differences of the measurements (i.e. $= SO_{2_{ijtl}} - HbO_{2_{ijtl}}$) for the i th donor at the j th level of foetal haemoglobin percent (Fhbperct) and the t th repeated measurement by the l th practitioner of the experiment.

(Carstensen (2004) also advocates the use of LME models in comparing methods, but with a different emphasis.)

7.5 Lai Shiao

Lai and Shiao (2005) use mixed models to determine the factors that affect the difference of two methods of measurement using the conventional formulation of linear mixed effects models.

If the parameter \mathbf{b} , and the variance components are not significantly different from zero, the conclusion that there is no inter-method bias can be drawn. If the fixed effects component contains only the intercept, and a simple correlation coefficient is used, then the estimate of the intercept in the model is the inter-method bias. Conversely the estimates for the fixed effects factors can advise the respective influences each factor has on the differences. The Proc Mixed package allows users to specify different correlation structures of the variance components \mathbf{G} and \mathbf{R} .

Oxygen saturation is one of the most frequently measured variables in clinical nursing studies. ‘Fractional saturation’ (HbO_2) is considered to be the gold standard method of measurement, with ‘functional saturation’ (SO_2) being an alternative method. The method of examining the causes of differences between these two methods is applied to a clinical study conducted by ?. This experiment was conducted by 8 lab practitioners on blood samples, with varying levels of haemoglobin, from two donors. The samples have been in storage for varying periods (described by the variable ‘Bloodage’) and are categorized according to haemoglobin percentages (i.e 0%,20%,40%,60%,80%,100%). There are 625 observations in all.

Lai and Shiao (2005) fits two models on this data, with the lab technicians and the replicate measurements as the random effects in both models. The first model uses haemoglobin level as a fixed effects component. For the second model, blood age is added as a second fixed factor.

Single fixed effect

The first model fitted by Lai and Shiao (2005) takes the blood level as the sole fixed effect to be analyzed. The following coefficient estimates are estimated by ‘Proc Mixed’;

$$\begin{aligned} \text{fixed effects : } & 2.5056 - 0.0263\text{Fhbperct}_{ijtl} & (7.1) \\ (\text{p-values : } & = 0.0054, < 0.0001, < 0.0001) \end{aligned}$$

$$\begin{aligned} \text{random effects : } & u(\sigma^2 = 3.1826) + e_{ijtl}(\sigma_e^2 = 0.1525, \rho = 0.6978) \\ (\text{p-values : } & = 0.8113, < 0.0001, < 0.0001) \end{aligned}$$

With the intercept estimate being both non-zero and statistically significant ($p = 0.0054$), this models supports the presence inter-method bias is 2.5% in favour of SO_2 . Also, the negative value of the haemoglobin level coefficient indicate that differences will decrease by 0.0263% for every percentage increase in the haemoglobin .

In the random effects estimates, the variance due to the practitioners is 3.1826, indicating that there is a significant variation due to technicians ($p = 0.0311$) affecting the differences. The variance for the estimates is given as 0.1525, ($p < 0.0001$).

Two fixed effects

Blood age is added as a second fixed factor to the model, whereupon new estimates are calculated;

$$\begin{aligned} \text{fixed effects : } & - 0.2866 + 0.1072\text{Bloodage}_{ijtl} - 0.0264\text{Fhbperct}_{ijtl} \\ (\text{p-values : } & = 0.8113, < 0.0001, < 0.0001) \end{aligned}$$

$$\begin{aligned} \text{random effects : } & u(\sigma^2 = 10.2346) + e_{ijtl}(\sigma_e^2 = 0.0920, \rho = 0.5577) \\ (\text{p-values : } & = 0.0446, < 0.0001, < 0.0001) & (7.2) \end{aligned}$$

With this extra fixed effect added to the model, the intercept term is no longer statistically significant. Therefore, with the presence of the second fixed factor, the

model is no longer supporting the presence of inter-method bias. Furthermore, the second coefficient indicates that the blood age of the observation has a significant bearing on the size of the difference between both methods ($p < 0.0001$). Longer storage times for blood will lead to higher levels of particular blood factors such as MetHb and HbCO (due to the breakdown and oxidisation of the haemoglobin). Increased levels of MetHb and HbCO are concluded to be the cause of the differences. The coefficient for the haemoglobin level doesn't differ greatly from the single fixed factor model, and has a much smaller effect on the differences. The random effects estimates also indicate significant variation for the various technicians; 10.2346 with $p = 0.0446$.

Lai and Shiao (2005) demonstrates how that linear mixed effects models can be used to provide greater insight into the cause of the differences. Naturally the addition of further factors to the model provides for more insight into the behavior of the data.

7.6 Lai Shiao

Lai and Shiao (2005) use mixed models to determine the factors that affect the difference of two methods of measurement using the conventional formulation of linear mixed effects models.

If the parameter \mathbf{b} , and the variance components are not significantly different from zero, the conclusion that there is no inter-method bias can be drawn. If the fixed effects component contains only the intercept, and a simple correlation coefficient is used, then the estimate of the intercept in the model is the inter-method bias. Conversely the estimates for the fixed effects factors can advise the respective influences each factor has on the differences. It is possible to pre-specify different correlation structures of the variance components \mathbf{G} and \mathbf{R} .

Oxygen saturation is one of the most frequently measured variables in clinical nursing studies. 'Fractional saturation' (HbO_2) is considered to be the gold standard method of measurement, with 'functional saturation' (SO_2) being an alternative method. The method of examining the causes of differences between these two meth-

ods is applied to a clinical study conducted by ?. This experiment was conducted by 8 lab practitioners on blood samples, with varying levels of haemoglobin, from two donors. The samples have been in storage for varying periods (described by the variable ‘Bloodage’) and are categorized according to haemoglobin percentages (i.e 0%,20%,40%,60%,80%,100%). There are 625 observations in all.

Lai and Shiao (2005) fits two models on this data, with the lab technicians and the replicate measurements as the random effects in both models. The first model uses haemoglobin level as a fixed effects component. For the second model, blood age is added as a second fixed factor.

Single fixed effect

The first model fitted by Lai and Shiao (2005) takes the blood level as the sole fixed effect to be analyzed. The following coefficient estimates are estimated by ‘Proc Mixed’;

$$\text{fixed effects : } 2.5056 - 0.0263\text{Fhbperct}_{ijtl} \quad (7.3)$$

$$(\text{p-values : } = 0.0054, < 0.0001, < 0.0001)$$

$$\text{random effects : } u(\sigma^2 = 3.1826) + e_{ijtl}(\sigma_e^2 = 0.1525, \rho = 0.6978)$$

$$(\text{p-values : } = 0.8113, < 0.0001, < 0.0001)$$

With the intercept estimate being both non-zero and statistically significant ($p = 0.0054$), this models supports the presence inter-method bias is 2.5% in favour of SO_2 . Also, the negative value of the haemoglobin level coefficient indicate that differences will decrease by 0.0263% for every percentage increase in the haemoglobin .

In the random effects estimates, the variance due to the practitioners is 3.1826, indicating that there is a significant variation due to technicians ($p = 0.0311$) affecting the differences. The variance for the estimates is given as 0.1525, ($p < 0.0001$).

Two fixed effects

Blood age is added as a second fixed factor to the model, whereupon new estimates are calculated;

$$\begin{aligned} \text{fixed effects : } & -0.2866 + 0.1072\text{Bloodage}_{ijtl} - 0.0264\text{Fhbperct}_{ijtl} \\ & (\text{p-values : } = 0.8113, < 0.0001, < 0.0001) \end{aligned}$$

$$\begin{aligned} \text{random effects : } & u(\sigma^2 = 10.2346) + e_{ijtl}(\sigma_e^2 = 0.0920, \rho = 0.5577) \\ & (\text{p-values : } = 0.0446, < 0.0001, < 0.0001) \end{aligned} \quad (7.4)$$

With this extra fixed effect added to the model, the intercept term is no longer statistically significant. Therefore, with the presence of the second fixed factor, the model is no longer supporting the presence of inter-method bias. Furthermore, the second coefficient indicates that the blood age of the observation has a significant bearing on the size of the difference between both methods ($p < 0.0001$). Longer storage times for blood will lead to higher levels of particular blood factors such as MetHb and HbCO (due to the breakdown and oxidation of the haemoglobin). Increased levels of MetHb and HbCO are concluded to be the cause of the differences. The coefficient for the haemoglobin level doesn't differ greatly from the single fixed factor model, and has a much smaller effect on the differences. The random effects estimates also indicate significant variation for the various technicians; 10.2346 with $p = 0.0446$.

Lai and Shiao (2005) demonstrates how that linear mixed effects models can be used to provide greater insight into the cause of the differences. Naturally the addition of further factors to the model provides for more insight into the behavior of the data.

7.7 Lai Shiao

Lai and Shiao (2005) use mixed models to determine the factors that affect the difference of two methods of measurement using the conventional formulation of linear mixed effects models.

If the parameter \mathbf{b} , and the variance components are not significantly different from zero, the conclusion that there is no inter-method bias can be drawn. If the fixed effects component contains only the intercept, and a simple correlation coefficient is used, then the estimate of the intercept in the model is the inter-method bias. Conversely the estimates for the fixed effects factors can advise the respective influences each factor has on the differences. It is possible to pre-specify different correlation structures of the variance components \mathbf{G} and \mathbf{R} .

Oxygen saturation is one of the most frequently measured variables in clinical nursing studies. ‘Fractional saturation’ (HbO_2) is considered to be the gold standard method of measurement, with ‘functional saturation’ (SO_2) being an alternative method. The method of examining the causes of differences between these two methods is applied to a clinical study conducted by ?. This experiment was conducted by 8 lab practitioners on blood samples, with varying levels of haemoglobin, from two donors. The samples have been in storage for varying periods (described by the variable ‘Bloodage’) and are categorized according to haemoglobin percentages (i.e 0%,20%,40%,60%,80%,100%). There are 625 observations in all.

Lai and Shiao (2005) fits two models on this data, with the lab technicians and the replicate measurements as the random effects in both models. The first model uses haemoglobin level as a fixed effects component. For the second model, blood age is added as a second fixed factor.

Single fixed effect

The first model fitted by Lai and Shiao (2005) takes the blood level as the sole fixed effect to be analyzed. The following coefficient estimates are estimated by ‘Proc Mixed’;

$$\begin{aligned} \text{fixed effects : } & 2.5056 - 0.0263\text{Fhbperct}_{ijtl} & (7.5) \\ (\text{p-values : } & = 0.0054, < 0.0001, < 0.0001) \end{aligned}$$

$$\begin{aligned} \text{random effects : } & u(\sigma^2 = 3.1826) + e_{ijtl}(\sigma_e^2 = 0.1525, \rho = 0.6978) \\ (\text{p-values : } & = 0.8113, < 0.0001, < 0.0001) \end{aligned}$$

With the intercept estimate being both non-zero and statistically significant ($p = 0.0054$), this models supports the presence inter-method bias is 2.5% in favour of SO_2 . Also, the negative value of the haemoglobin level coefficient indicate that differences will decrease by 0.0263% for every percentage increase in the haemoglobin .

In the random effects estimates, the variance due to the practitioners is 3.1826, indicating that there is a significant variation due to technicians ($p = 0.0311$) affecting the differences. The variance for the estimates is given as 0.1525, ($p < 0.0001$).

Two fixed effects

Blood age is added as a second fixed factor to the model, whereupon new estimates are calculated;

$$\begin{aligned} \text{fixed effects : } & -0.2866 + 0.1072\text{Bloodage}_{ijtl} - 0.0264\text{Fhbperct}_{ijtl} \\ (\text{p-values : } & = 0.8113, < 0.0001, < 0.0001) \\ \\ \text{random effects : } & u(\sigma^2 = 10.2346) + e_{ijtl}(\sigma_e^2 = 0.0920, \rho = 0.5577) \\ (\text{p-values : } & = 0.0446, < 0.0001, < 0.0001) & (7.6) \end{aligned}$$

With this extra fixed effect added to the model, the intercept term is no longer statistically significant. Therefore, with the presence of the second fixed factor, the

model is no longer supporting the presence of inter-method bias. Furthermore, the second coefficient indicates that the blood age of the observation has a significant bearing on the size of the difference between both methods ($p < 0.0001$). Longer storage times for blood will lead to higher levels of particular blood factors such as MetHb and HbCO (due to the breakdown and oxidisation of the haemoglobin). Increased levels of MetHb and HbCO are concluded to be the cause of the differences. The coefficient for the haemoglobin level doesn't differ greatly from the single fixed factor model, and has a much smaller effect on the differences. The random effects estimates also indicate significant variation for the various technicians; 10.2346 with $p = 0.0446$.

Lai and Shiao (2005) demonstrates how that linear mixed effects models can be used to provide greater insight into the cause of the differences. Naturally the addition of further factors to the model provides for more insight into the behavior of the data.

7.8 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 Bland and Altman (1999). 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 com-

pound symmetry (CS) form, and the alternative model in symmetric (symm) form. These matrices are determined to be as follows;

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

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.

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

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;

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

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 Hamlett et al. (2004).

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

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 .

7.9 Worked Examples

7.9.1 LikelihoodRatio Tests

Conventionally LME models can be tested using Likelihood Ratio Tests, wherein a reference model is compared to a nested model.

```

> Ref.Fit = lme(y ~ meth-1, data = dat,    #Symm , Symm#
+   random = list(item=pdSymm(~ meth-1)),
+   weights=varIdent(form=~1|meth),
+   correlation = corSymm(form=~1 | item/repl),
+   method="ML")

```

Roy(2009) presents two nested models that specify the condition of equality as required, with a third nested model for an additional test. There three formulations share the same structure, and can be specified by making slight alterations of the code for the Reference Model. Nested Model (Between-Item Variability)

```

> NMB.fit = lme(y ~ meth-1, data = dat,    #CS , Symm#
+   random = list(item=pdCompSymm(~ meth-1)),
+   correlation = corSymm(form=~1 | item/repl),
+   method="ML")

```

Nested Model (Within item Variability)

```

> NMW.fit = lme(y ~ meth-1, data = dat,    #Symm , CS#
+   random = list(item=pdSymm(~ meth-1)),
+   weights=varIdent(form=~1|meth),
+   correlation = corCompSymm(form=~1 | item/repl),
+   method="ML")

```

Nested Model (Overall Variability) Additionally there is a third nested model, that can be used to test overall variability, substantively a a joint test for between-item

and within-item variability. The motivation for including such a test in the suite is not clear, although it does circumvent the need for multiple comparison procedures in certain circumstances, hence providing a simplified procedure for non-statisticians.

```
> NMO.fit = lme(y ~ meth-1, data = dat,    #CS , CS#
+   random = list(item=pdCompSymm(~ meth-1)),
+   correlation = corCompSymm(form=~1 | item/repl),
+   method="ML")
```

ANOVAs for Original Fits The likelihood Ratio test is very simple to implement in R. All that is required is to specify the reference model and the relevant nested mode as arguments to the command `anova()`. The figure below displays the three tests described by Roy (2009).

```
> testB    = anova(Ref.Fit,NMB.fit)                # Between-Subject Variability
> testW    = anova(Ref.Fit,NMW.fit)                # Within-Subject Variability
> testO    = anova(Ref.Fit,NMO.fit)                # Overall Variability
```

7.9.2 Reference Model (Ref.Fit)

Conventionally LME models can be tested using Likelihood Ratio Tests, wherein a reference model is compared to a nested model.

```
> Ref.Fit = lme(y ~ meth-1, data = dat, #Symm , Symm#  
+ random = list(item=pdSymm(~ meth-1)),  
+ weights=varIdent(form=~1|meth),  
+ correlation = corSymm(form=~1 | item/repl),  
+ method="ML")
```

Roy(2009) presents two nested models that specify the condition of equality as required, with a third nested model for an additional test. These three formulations share the same structure, and can be specified by making slight alterations of the code for the Reference Model.

7.9.3 Nested Model (Between-Item Variability)

```
> NMB.fit = lme(y ~ meth-1, data = dat, #CS , Symm#  
+ random = list(item=pdCompSymm(~ meth-1)),  
+ correlation = corSymm(form=~1 | item/repl),  
+ method="ML")
```

7.10 Testing Procedures

Roy's methodology requires the construction of four candidate models. The first candidate model is compared to each of the three other models successively. It is the

alternative model in each of the three tests, with the other three models acting as the respective null models.

The probability distribution of the test statistic can be approximated by a chi-square distribution with $(\nu_1 - \nu_2)$ degrees of freedom, where ν_1 and ν_2 are the degrees of freedom of models 1 and 2 respectively.

Likelihood ratio tests are very simple to implement in R, simply use the 'anova()' commands. Sample output will be given for each variability test. The likelihood ratio test is the procedure used to compare the fit of two models. For each candidate model, the '-2 log likelihood' ($M2LL$) is computed. The test statistic for each of the three hypothesis tests is the difference of the $M2LL$ for each pair of models. If the p -value in each of the respective tests exceed as significance level chosen by the analyst, then the null model must be rejected.

$$-2 \ln \Lambda_d = [M2LL \text{ under } H_0 \text{ model}] - [M2LL \text{ under } H_A \text{ model}] \quad (7.7)$$

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

$$\nu = [\text{LRT df under } H_0 \text{ model}] - [\text{LRT df under } H_A \text{ model}] \quad (7.8)$$

```
> 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
```

7.10.1 Roy's Reference Model

Conventionally LME models can be tested using Likelihood Ratio Tests, wherein a reference model is compared to a nested model.

```
> Ref.Fit = lme(y ~ meth-1, data = dat, #Symm , Symm#
+ random = list(item=pdSymm(~ meth-1)),
+ weights=varIdent(form=~1|meth),
+ correlation = corSymm(form=~1 | item/repl),
+ method="ML")
```

Roy(2009) presents two nested models that specify the condition of equality as required, with a third nested model for an additional test. There three formulations share the same structure, and can be specified by making slight alterations of the code for the Reference Model.

7.10.2 Nested Model (Between-Item Variability)

```
> NMB.fit = lme(y ~ meth-1, data = dat, #CS , Symm#
+ random = list(item=pdCompSymm(~ meth-1)),
+ correlation = corSymm(form=~1 | item/repl),
```



```
+ method="ML")
```

7.11 Worked Examples

Roy (2009b) uses examples from Bland and Altman (1986) to be able to compare both types of analysis.

? uses the “Blood” data set, which featured in Bland and Altman (1999).

7.12 Examples: LoAs for Carstensen’s data

Carstensen et al. (2008) 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,

$$0.0449 \pm 1.96 \times \sqrt{2 \times 0.0596^2 + 0.0772^2 + 0.0724^2} = (-0.220, 0.309). \quad (7.9)$$

For Carstensen’s ‘fat’ data, the limits of agreement computed using Roy’s method are consistent with the estimates given by Carstensen et al. (2008); $0.044884 \pm 1.96 \times 0.1373979 = (-0.224, 0.314)$.

? describes the calculation of the limits of agreement (with the inter-method bias implicit) for both data sets, based on his formulation;

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

For the ‘Fat’ data set, the inter-method bias is shown to be 0.045. The limits of agreement are $(-0.23, 0.32)$

Carstensen demonstrates the use of the interaction term when computing the limits of agreement for the ‘Oximetry’ data set. When the interaction term is omitted, the limits of agreement are $(-9.97, 14.81)$. Carstensen advises the inclusion of the

interaction term for linked replicates, and hence the limits of agreement are recomputed as $(-12.18, 17.12)$.

7.12.1 Diabetes Example

Carstensen et al. (2008) describes the sampling method when discussing of a motivating example

Diabetes patients attending an outpatient clinic in Denmark have their HbA_{1c} levels routinely measured at every visit. Venous and Capillary blood samples were obtained from all patients appearing at the clinic over two days. Samples were measured on four consecutive days on each machines, hence there are five analysis days.

Carstensen et al. (2008) notes that every machine was calibrated every day to the manufacturers guidelines. Measurements are classified by method, individual and replicate. In this case the replicates are clearly not exchangeable, neither within patients nor simulataneously for all patients.

7.12.2 Oximetry Data

Carstensen et al. (2008) 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. Carstensen et al. (2008) 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.

Carstensen et al. (2008) 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 additional 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);

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

Akaike (1974) introduces the Akaike information criterion (AIC), a model selection tool based on the likelihood function. Given a data set, candidate models are ranked according to their AIC values, with the model having the lowest AIC being considered the best fit. Two candidate models can be 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 $\hat{\mathbf{A}}$ 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 $\hat{\mathbf{A}}$. Therefore the test's proposed by Roy (2009b) 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 $15.62 \pm 1.96 \times 20.33 = (-24.22, 55.46)$.)

Carstensen demonstrates the use of the interaction term when computing the limits of agreement for the ‘Oximetry’ data set. When the interaction term is omitted, the limits of agreement are $(-9.97, 14.81)$. Carstensen advises the inclusion of the interaction term for linked replicates, and hence the limits of agreement are recomputed as $(-12.18, 17.12)$.

For Carstensen’s ‘fat’ data, the limits of agreement computed using Roy’s method are consistent with the estimates given by Carstensen et al. (2008); $0.044884 \pm 1.96 \times 0.1373979 = (-0.224, 0.314)$.

7.12.3 Linked replicates

Carstensen et al. (2008) 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.

Carstensen et al. (2008) 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. Carstensen et al. (2008) 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.

Carstensen et al. (2008) 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);

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

Akaike (1974) introduces the Akaike information criterion (AIC), a model selection tool based on the likelihood function. Given a data set, candidate models are ranked according to their AIC values, with the model having the lowest AIC being considered the best fit. Two candidate models can be 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 $\hat{\mathbf{A}}$ 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 $\hat{\mathbf{A}}$. Therefore the test's proposed by Roy (2009b) 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 $15.62 \pm 1.96 \times 20.33 = (-24.22, 55.46)$.

7.12.4 Limits of agreement for Carstensen’s data

? describes the calculation of the limits of agreement (with the inter-method bias implicit) for both data sets, based on his formulation;

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

For the ‘Fat’ data set, the inter-method bias is shown to be 0.045. The limits of agreement are $(-0.23, 0.32)$

Carstensen demonstrates the use of the interaction term when computing the limits of agreement for the ‘Oximetry’ data set. When the interaction term is omitted, the limits of agreement are $(-9.97, 14.81)$. Carstensen advises the inclusion of the interaction term for linked replicates, and hence the limits of agreement are recomputed as $(-12.18, 17.12)$.

7.12.5 Fat Data Examples: LoAs for Carstensen’s data

Carstensen et al. (2008) 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,

$$0.0449 \pm 1.96 \times \sqrt{2 \times 0.0596^2 + 0.0772^2 + 0.0724^2} = (-0.220, 0.309). \quad (7.10)$$

Carstensen et al. (2008) describes the calculation of the limits of agreement (with the inter-method bias implicit) for both data sets, based on his formulation;

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

For the ‘Fat’ data set, the inter-method bias is shown to be 0.045. The limits of agreement are $(-0.23, 0.32)$

For Carstensen’s ‘fat’ data, the limits of agreement computed using Roy’s method are consistent with the estimates given by Carstensen et al. (2008); $0.044884 \pm 1.96 \times 0.1373979 = (-0.224, 0.314)$.

7.12.6 RV-IV

For the the RV-IC comparison, \hat{D} is given by

$$\hat{D} = \begin{bmatrix} 1.6323 & 1.1427 \\ 1.1427 & 1.4498 \end{bmatrix} \quad (7.11)$$

The estimate for the within-subject variance covariance matrix is given by

$$\hat{\Sigma} = \begin{bmatrix} 0.1072 & 0.0372 \\ 0.0372 & 0.1379 \end{bmatrix} \quad (7.12)$$

The estimated overall variance covariance matrix for the the ‘RV vs IC’ comparison is given by

$$Block\Omega_i = \begin{bmatrix} 1.7396 & 1.1799 \\ 1.1799 & 1.5877 \end{bmatrix}. \quad (7.13)$$

The power of the likelihood ratio test may depends on specific sample size and the specific number of replications, and the author proposes simulation studies to examine this further.

7.12.7 Limits of agreement for Oximetry

Carstensen demonstrates the use of the interaction term when computing the limits of agreement for the ‘Oximetry’ data set. When the interaction term is omitted, the limits of agreement are $(-9.97, 14.81)$. Carstensen advises the inclusion of the interaction term for linked replicates, and hence the limits of agreement are recomputed as $(-12.18, 17.12)$.

7.12.8 Classical Model

The classical model is based on measurements y_{mi} by method $m = 1, 2$ on item $i = 1, 2, \dots$

$$y_{mi} = \alpha_m + \mu_i + e_{mi}$$

$$e_{mi} \sim N(0, \sigma_m^2)$$

Even though the separate variances can not be identified, their sum can be estimated by the empirical variance of the differences.

Like wise the separate α can not be estimated, only their difference can be estimated as \bar{D}

7.13 Classical model for single measurements

The classical model is based on measurements y_{mi} by method $m = 1, 2$ on item $i = 1, 2, \dots$

$$y_{mi} = \alpha_m + \mu_i + e_{mi}$$

$$e_{mi} \sim N(0, \sigma_m^2)$$

Even though the separate variances can not be identified, their sum can be estimated by the empirical variance of the differences.

Like wise the separate α can not be estimated, only their difference can be estimated as \bar{D}

In the first instance, we require a simple model to describe a measurement by method m . We use the term *item* to denote an individual, subject or sample, to be measured, being randomly sampled from a population. Let y_{mi} be the measurement for item i made by method m .

$$y_{mi} = \alpha_m + \mu_i + e_{mi}$$

- α_m is the fixed effect associated with method m ,
- μ_i is the true value for subject i (fixed effect),
- e_{mi} is a random effect term for errors with $e_{mi} \sim \mathcal{N}(0, \sigma_m^2)$.

This model implies that the difference between the paired measurements can be expressed as

$$d_i = y_{1i} - y_{2i} \sim \mathcal{N}(\alpha_1 - \alpha_2, \sigma_1^2 - \sigma_2^2).$$

Importantly, this is independent of the item levels μ_i . As the case-wise differences are of interest, the parameters of interest are the fixed effects for methods α_m .

$$y_{mi} = \alpha_m + \mu_i + e_{mi}$$

Importantly these variance covariance structures are central to Roy methodology.

? proposes a series of hypothesis tests based on these matrices as part of her methodology. These tests shall be reverted to in due course.

The standard deviation of the differences of variables a and b is computed as

$$\text{var}(a - b) = \text{var}(a) + \text{var}(b) - 2\text{cov}(a, b)$$

Hence the variance of the difference of two methods, that allows for the calculation of the limits of agreement, can be calculated as

$$\text{var}(d) = \omega_1^2 + \omega_2^2 - 2 \times \omega_1 \omega_2$$

7.13.1 Sampling

One important feature of replicate observations is that they should be independent of each other. In essence, this is achieved by ensuring that the observer makes each measurement independent of knowledge of the previous value(s). This may be difficult to achieve in practice. (Check who said this)

7.13.2 Remarks on the Multivariate Normal Distribution

Diligence is required when considering the models. Carstensen specifies his models in terms of the univariate normal distribution. Roy's model is specified using the bivariate normal distribution. This gives rise to a key difference between the two models, in that a bivariate model accounts for covariance between the variables of interest. The multivariate normal distribution of a k -dimensional random vector $X = [X_1, X_2, \dots, X_k]$ can be written in the following notation:

$$X \sim \mathcal{N}(\mu, \Sigma),$$

or to make it explicitly known that X is k -dimensional,

$$X \sim \mathcal{N}_k(\mu, \Sigma).$$

with k -dimensional mean vector

$$\mu = [E[X_1], E[X_2], \dots, E[X_k]]$$

and $k \times k$ covariance matrix

$$\Sigma = [\text{Cov}[X_i, X_j]], \quad i = 1, 2, \dots, k; \quad j = 1, 2, \dots, k$$

1. Univariate Normal Distribution

$$X \sim \mathcal{N}(\mu, \sigma^2),$$

2. Bivariate Normal Distribution

(a)

$$X \sim \mathcal{N}_2(\mu, \Sigma),$$

(b)

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}.$$

Bibliography

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6), 716–723.
- Barnhart, H., M. Haber, and L. Lin (2007). An overview of assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics* 17, 529–569.
- Bland, J. and D. Altman (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* i, 307–310.
- Bland, J. and D. Altman (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 8(2), 135–160.
- Carstensen, B. (2004). Comparing and predicting between several methods of measurement. *Biostatistics* 5(3), 399–413.
- Carstensen, B. (2010). *Comparing Clinical Measurement Methods : A Practical Guide*. Wiley.
- Carstensen, B., J. Simpson, and L. C. Gurrin (2008). Statistical models for assessing agreement in method comparison studies with replicate measurements. *The International Journal of Biostatistics* 4(1).
- Cook, R. (1986). Assessment of local influence. *Journal of the Royal Statistical Society. Series B (Methodological)* 48(2), 133–169.
- Grubbs, F. (1948). On estimating precision of measuring instruments and product variability. *Journal of the American Statistical Association* 43, 243–264.

- Hamlett, A., L. Ryan, and R. Wolfinger (2004). On the use of PROC MIXED to estimate correlation in the presence of repeated measures. *Proceedings of the Statistics and Data Analysis Section, SAS Users Group International 198-229*, 1–7.
- Lai, D. and S.-Y. P. K. Shiao (2005). Comparing two clinical measurements: a linear mixed model approach. *Journal of Applied Statistics* 32(8), 855–860.
- Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics* 38(4), 963–974.
- Lam, M., K. Webb, and D. O’Donnell (1999). Correlation between two variables in repeated measurements. *American Statistical Association, Proceedings of the Biometric Session*, 213–218.
- Pinheiro, J. and D. Bates (1994). *Mixed Effects Models in S and S plus* (2nd ed.). Reading, Massachusetts: Springer.
- Roy, A. (2006). Estimating correlation coefficient between two variables with repeated observations using mixed effects models. *Biometric Journal* 2, 286–301.
- Roy, A. (2009a). An application of linear mixed effects model to assess the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics* 19, 150–173.
- Roy, A. (2009b). An application of the linear mixed effects model to ass the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics* 19, 150–173.
- West, B., K. Welch, and A. Galecki (2007). *Linear Mixed Models: a Practical Guide Using Statistical Software*. Chapman and Hall CRC.

7.14 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 Roy (2009a), 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.

7.15 Conclusion

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