Contents

Chapter 1

Introduction

1.1 Repeated Measurements

In cases where there are repeated measurements by each of the two methods on the same subjects , Bland Altman suggest calculating the mean for each method on each subject and use these pairs of means to compare the two methods. The estimate of bias will be unaffected using this approach, but the estimate of the standard deviation of the differences will be too small, because of the reduction of the effect of repeated measurement error. Bland Altman propose a correction for this. Carstensen attends to this issue also, adding that another approach would be to treat each repeated measurement separately.

1.2 Replicate measurements

Roy 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.

1.3 Linear Mixed effects Models

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

effects are considered non-random and the levels of each factor are of specific interest.

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

The methodology has developed since, including contributions from **?**, who extend the use of variance components into linear models, and **?**, who introduced the 'mixed model' terminology and formally distinguished between mixed and random effects models. **?** devised a methodology for deriving estimates for both the fixed effects and the random effects, using a set of equations that would become known as 'mixed model equations' or 'Henderson's equations'. LME methodology is further enhanced by Henderson's later works (**?????**). The key features of Henderson's work shall be discussed in depth in due course.

? demonstrated that unique estimates of the variance components could be obtained using maximum likelihood methods. These estimates would later become known as the maximum likelihood (ML) estimates. However these estimates are known to be biased downwards, because of the assumption that the fixed estimates are known, rather than being estimated from the data. **?** produced an alternative set of estimates, known as the restricted maximum likelihood (REML) estimates, that do not require the fixed effects to be known. Both ML and REML estimation shall be reverted to in due course.

? provides a form of notation for notation for LME models that has since become the standard form, or the basis for more complex formulations. Due to computation complexity, linear mixed effects models have not seen widespread use until many well known statistical software applications began facilitating them. SAS Institute added PROC MIXED to its software suite in 1992 (**?**). **?** described how to compute LME models in the `R` environment.

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

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

where y is a vector of N observable random variables, β is a vector of p fixed effects,

X and Z are $N \times p$ and $N \times q$ known matrices, and b and ϵ are vectors of q and N, respectively, random effects such that $\mathrm{E}(b) = 0,\ \mathrm{E}(\epsilon) = 0$ and

```
\[
\mathrm{var}
\pmatrix{
b \cr
\epsilon }  =
\pmatrix{
D & 0 \cr
0 & \Sigma }
\]
```

where D and Σ are positive definite matrices parameterized by an unknown variance component parameter vector θ. The notation provided here is generic, and will be adapted to accord with complex formulations that will be encountered in due course.

1.3.1 Likelihood and estimation

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

The likelihood function ($L(\theta)$)is a fundamental concept in statistical inference. It indicates how likely a particular population is to produce an observed sample. The set of values that maximize the likelihood function are considered to be optimal, and are used as the estimates of the parameters. For computational ease, it is common to use the logarithm of the likelihood function, known simply as the log-likelihood ($\ell(\theta)$).

Likelihood-based tools

Likelihood functions provide the basis for two important statistical concepts that shall be further referred to; the likelihood ratio test and the Akaike information criterion.

Likelihood 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.

? 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.

Likelihood estimation techniques

Maximum likelihood and restricted maximum likelihood have become the most common strategies for estimating the variance component parameter θ. Maximum likelihood estimation obtains parameter estimates by optimizing the likelihood function. To obtain ML estimate the likelihood is constructed as a function of the parameters in the specified LME model. The maximum likelihood estimates (MLEs) of the parameters are the values of the arguments that maximize the likelihood function. The REML approach is a variant of maximum likelihood estimation which does not base estimates on a maximum likelihood fit of all the information, but instead uses a likelihood function derived from a data set, transformed to remove the irrelevant influences (**?**).

Restricted maximum likelihood is often preferred to maximum likelihood because REML estimation reduces the bias in the variance component by taking into account the loss of degrees of freedom that results from estimating the fixed effects in $\boldsymbol{\beta}$. Restricted maximum likelihood also handles high correlations more effectively, and is less sensitive to outliers than maximum likelihood. The problem with REML for model building is that the likelihoods obtained for different fixed effects are not comparable. Hence it is not valid to compare models with different fixed effects using a likelihood

ratio test or AIC when REML is used to estimate the model. Therefore models derived using ML must be used instead.

1.3.2 Henderson's equations

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

```
\begin{equation}
\left|
\pmatrix{
D & 0 \cr
0 & \Sigma }
\right|^{-\frac{1}{2}}
\exp
\left\{ -\frac{1}{2}
\pmatrix{
b \cr
y - X \beta -Zb
}^\prime
\pmatrix{
D & 0 \cr
0 & \Sigma }^{-1}
\pmatrix{
b \cr
y - X \beta -Zb
}
\right\},
\label{u&beta:JointDensity}
\end{equation}
```

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

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

This leads to the mixed model equations

```
\begin{equation}
\pmatrix{
X^\prime\Sigma^{-1}X & X^\prime\Sigma^{-1}Z
\cr
Z^\prime\Sigma^{-1}X & X^\prime\Sigma^{-1}X + D^{-1}
}
\pmatrix{
\beta \cr
b
}
=
\pmatrix{
X^\prime\Sigma^{-1}y \cr
Z^\prime\Sigma^{-1}y
}.
\label{Henderson:Equations}
\end{equation}
```

? points out that although **?** initially referred to the estimates $\hat{\beta}$ and $\hat{b}$ from (1.8) as "joint maximum likelihood estimates". **?** later advised that these estimates should not be referred to as "maximum likelihood" as the function being maximized in (1.7) is a joint density rather than a likelihood function.

Estimation of the fixed parameters

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

eters (β, θ) is

$$\ell(\beta, \theta|y) = -\frac{1}{2}\log|V| - \frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta), \tag{1.3}$$

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

$$(X'V^{-1}X)\beta = X'V^{-1}y. \tag{1.4}$$

Substituting $\hat{\beta}$ from (1.10) into $\ell(\beta, \theta|y)$ from (1.9) returns the *profile* log-likelihood

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

of the variance parameter θ. Estimates of the parameters θ specifying V can be found by maximizing $\ell_P(\theta \mid y)$ over θ. These are the ML estimates. For REML estimation the *restricted* log-likelihood is defined as

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

Estimation of the random effects

The established approach for estimating the random effects is to use the best linear predictor of b from y, which for a given β equals $DZ'V^{-1}(y - X\beta)$. In practice β is replaced by an estimator such as $\hat{\beta}$ from (1.10) so that $\hat{b} = DZ'V^{-1}(y - X\hat{\beta})$. Pre-multiplying by the appropriate matrices it is straightforward to show that these estimates $\hat{\beta}$ and $\hat{b}$ satisfy the equations in (1.8).

1.3.3 Algorithms for likelihood function optimization

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

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

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

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

1.3.4 The extended likelihood

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

and adopting the same distributional assumptions used by **?** yields the log-likelihood function

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

Given θ, differentiating with respect to β and b returns Henderson's equations in (1.8).

Henderson's equations in (1.8) can be rewritten $(T'W^{-1}T)\delta = T'W^{-1}y_a$ using

```
\[
\delta = \pmatrix{\beta \cr b},
\ y_{a} = \pmatrix{
y \cr \psi
},
\ T = \pmatrix{
X & Z  \cr
0 & I
},
\ \textrm{and} \ W = \pmatrix{
\Sigma & 0  \cr
0 &  D },
\]
```

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

1.4 Linear Mixed effects Models

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

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

The methodology has developed since, including contributions from **?**, who extend the use of variance components into linear models, and **?**, who introduced the 'mixed model' terminology and formally distinguished between mixed and random effects models. **?** devised a methodology for deriving estimates for both the fixed effects and the random effects, using a set of equations that would become known as 'mixed model equations' or 'Henderson's equations'. LME methodology is further enhanced by Henderson's later works (**?????**). The key features of Henderson's work provide the basis for the estimation techniques.

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

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

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

$$y = X\beta + Zb + \epsilon \tag{1.5}$$

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

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

where D and Σ are positive definite matrices parameterized by an unknown variance component parameter vector θ. The variance-covariance matrix for the vector of observations y is given by $V = ZDZ' + \Sigma$. This implies $y \sim (X\beta, V) = (X\beta, ZDZ' + \Sigma)$. It is worth noting that V is an $n \times n$ matrix, as the dimensionality becomes relevant later on. The notation provided here is generic, and will be adapted to accord with complex formulations that will be encountered in due course.

1.4.1 Why use LMEs?

The LME model approach has seen increased use as a framework for method comparison studies in recent years (Lai & Shaio, Carstensen and Choudhary as examples). In part this is due to the increased profile of LME models, and furthermore the availability of capable software. Additionally LME based approaches may utilise the diagnostic and influence analysis techniques that have been developed in recent times.

1.4.2 Estimation

Estimation of LME models involve two complementary estimation issues'; estimating the vectors of the fixed and random effects estimates $\hat{\beta}$ and $\hat{b}$ and estimating the variance covariance matrices D and Σ. Inference about fixed effects have become known as 'estimates', while inferences about random effects have become known as 'predictions'. The most common approach to obtain estimators are Best Linear Unbiased Estimator (BLUE) and Best Linear Unbiased Predictor (BLUP). For an LME model given by (1.5), the BLUE of $\hat{\beta}$ is given by

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

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

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

Henderson's equations

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

$$\left| \begin{array}{cc} D & 0 \\ 0 & \Sigma \end{array} \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix}' \begin{pmatrix} D & 0 \\ 0 & \Sigma \end{pmatrix}^{-1} \begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix} \right\}, \quad (1.6)$$

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

$$(y - X\beta - Zb)'\Sigma^{-1}(y - X\beta - Zb) + b'D^{-1}b. \quad (1.7)$$

This leads to the mixed model equations

$$\begin{pmatrix} X'\Sigma^{-1}X & X'\Sigma^{-1}Z \\ Z'\Sigma^{-1}X & X'\Sigma^{-1}X + D^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X'\Sigma^{-1}y \\ Z'\Sigma^{-1}y \end{pmatrix}. \quad (1.8)$$

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

? points out that although **?** initially referred to the estimates $\hat{\beta}$ and $\hat{b}$ from (1.8) as "joint maximum likelihood estimates", **?** later advised that these estimates should not be referred to as "maximum likelihood" as the function being maximized in (1.7) is a joint density rather than a likelihood function. **?** remarks that it is clear that Henderson used joint estimation for computational purposes, without recognizing the theoretical implications.

Estimation of the fixed parameters

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

$$\ell(\beta, \theta|y) = -\frac{1}{2}\log|V| - \frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta), \tag{1.9}$$

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

$$(X'V^{-1}X)\beta = X'V^{-1}y. \tag{1.10}$$

Substituting $\hat{\beta}$ from (1.10) into $\ell(\beta, \theta|y)$ from (1.9) returns the *profile* log-likelihood

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

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

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

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

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

Estimation of the random effects

The established approach for estimating the random effects is to use the best linear predictor of b from y, which for a given β equals $DZ'V^{-1}(y - X\beta)$. In practice β is replaced by an estimator such as $\hat{\beta}$ from (1.10) so that $\hat{b} = DZ'V^{-1}(y - X\hat{\beta})$. Pre-multiplying by the appropriate matrices it is straightforward to show that these estimates $\hat{\beta}$ and $\hat{b}$ satisfy the equations in (1.8).

Algorithms for likelihood function optimization

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

The EM algorithm, introduced by **?**, is an iterative technique for maximizing complicated likelihood functions. The algorithm alternates between performing an expectation (E) step and the maximization (M) step. The 'E' step computes the expectation

of the log-likelihood evaluated using the current estimate for the variables. In the 'M' step, parameters that maximize the expected log-likelihood, found on the previous 'E' step, are computed. These parameter estimates are then used to determine the distribution of the variables in the next 'E' step. The algorithm alternatives between these two steps until convergence is reached.

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

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

The extended likelihood

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

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

Given θ, differentiating with respect to β and b returns Henderson's equations in (1.8).

1.5 Statement of the LME model

Further to a paper published by Laird and Ware in 1982, a linear mixed effects model is a linear mdoel that combined fixed and random effect terms formulated as follows;

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

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

1.6 Repeated measurements in LME models

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

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

1.6.1 Statistical model

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

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

$$\begin{pmatrix} y_{Aij} \\ y_{Bij} \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ where } \boldsymbol{\mu} = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \tag{1.11}$$

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

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix} \tag{1.12}$$

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

1.6.2 Formulation of the response vector

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

Consider the case where three measurements are taken by both methods A and B, $\boldsymbol{y_i}$ is a 6×1 random vector describing the ith subject.

$$\boldsymbol{y}_i = (y_i^{A1}, y_i^{B1}, y_i^{A2}, y_i^{B2}, y_i^{A3}, y_i^{B3})\boldsymbol{\prime} \tag{1.13}$$

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

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{b_i} + \boldsymbol{\epsilon_i} \tag{1.14}$$

$$\boldsymbol{b_i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}) \tag{1.15}$$

$$\boldsymbol{\epsilon_i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R_i}) \tag{1.16}$$

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

1.6.3 Decomposition of the response covariance matrix

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

$$\mathrm{Cov}(\mathrm{y}_i) = \boldsymbol{\Omega_i} = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i\boldsymbol{\prime} + \boldsymbol{R_i}. \tag{1.17}$$

$\boldsymbol{R_i}$ can be shown to be the Kronecker product of a correlation matrix $\boldsymbol{V}$ and $\boldsymbol{\Lambda}$. The correlation matrix $\boldsymbol{V}$ of the repeated measures on a given response variable is assumed to be the same for all response variables. **?** 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. Various error structures may be assumed for repeated measure correlations. However both **?** and **?** use the identity matrix, with dimensions $n_i \times n_i$ as the formulation for $\boldsymbol{V}$. This has the benefit of being simple to implement. Roy's results also indicate the use of the identity matrix.

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

$$\boldsymbol{\Omega}_i = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{D} & \boldsymbol{D} \\ \boldsymbol{D} & \boldsymbol{\Sigma} & \boldsymbol{D} \\ \boldsymbol{D} & \boldsymbol{D} & \boldsymbol{\Sigma} \end{pmatrix} \tag{1.18}$$

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

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

$$\boldsymbol{\Omega_i} = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_{i\prime} + (\boldsymbol{I}_{\boldsymbol{n_i}} \otimes \boldsymbol{\Lambda}). \tag{1.19}$$

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

1.6.4 Correlation terms

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

$$\boldsymbol{D} = \begin{pmatrix} \sigma_A^2 \rho_A & \sigma_A \sigma_b \rho_{AB} \delta \\ \sigma_A \sigma_b \rho_{AB} \delta & \sigma_B^2 \rho_B \end{pmatrix} \tag{1.20}$$

$$\boldsymbol{\Lambda} = \begin{pmatrix} \sigma_A^2 (1 - \rho_A) & \sigma_{AB} (1 - \delta) \\ \sigma_{AB} (1 - \delta) & \sigma_B^2 (1 - \rho_B) \end{pmatrix}. \tag{1.21}$$

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. **?** provides a useful graphical depiction of the role of each correlation coefficients.

1.7 Conclusion

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

1.8 LME models in method comparison studies

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

? further 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.

? views the uses of linear mixed effects models as an expansion on the Bland-Altman methodology, rather than as a replacement.**?** 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. **?** is interesting in that it extends the usual method comparison study question. It correctly identifies LME models as a methodoloy that can used to make such questions tractable.

? extends the usual method comparison study question. It correctly identifies LME models as a methodoloy that can used to make such questions tractable. The Data 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. **?** 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 adchieved thanks to authors such as **?**, **?**, **? ?**, amongst others. In this chapter various LME approaches to method comparison studies shall be examined.

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

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

1.9 The Linear Mixed Effects Model

The linear mixed effects model is given by

$$Y = X\beta + Zu + \epsilon \tag{1.22}$$

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

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

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

$$\mathbf{V} = var(\mathbf{Xb} + \mathbf{Zu} + \mathbf{e}) \tag{1.23}$$

$$\mathbf{V} = var(\mathbf{Xb}) + var(\mathbf{Zu}) + var(\mathbf{e})) \tag{1.24}$$

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

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

$$\mathbf{V} = ZGZ^T + \mathbf{R} \tag{1.25}$$

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

$$\hat{b} = (X^T V^{-1} X)^{-1} X^T V^{-1} y \tag{1.26}$$

$$\hat{u} = G Z^T V^{-1} (y - X\hat{b}) \tag{1.27}$$

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

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

1.9.1 Why use LMEs for Method Comparison?

The LME model approach has seen increased use as a framework for method comparison studies in recent years (Lai & Shaio, Carstensen and Choudhary as examples). In part this is due to the increased profile of LME models, and furthermore the availability of capable software. Additionally LME based approaches may utilise the diagnostic and influence analysis techniques that have been developed in recent times.

Roy proposes an LME model with Kronecker product covariance structure in a doubly multivariate setup. Response for ith subject can be written as

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

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

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

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

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

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

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

1.10 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.

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

1.11 Carstensen's Model

? 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;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2) \tag{1.28}$$

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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \tag{1.29}$$

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.

1.12 Two Way ANOVA

? 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

$$y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + \varepsilon_{mir}. \tag{1.30}$$

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 $m = \{1, 2, 3 \ldots\}$. Based on the model expressed in (2), **?** compute the limits of agreement as

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

? 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 (1.31) requires $N + 2$ fixed effects.

Allocating fixed effects to each item i by (1.31) accords with earlier work on comparing methods of measurement, such as **?**. However allocation of fixed effects in ANOVA models suggests that the group of items is itself of particular interest, rather than as a representative sample used of the overall population. However this approach seems contrary to the purpose of LOAs as a prediction interval for a population of items. Conversely, **?** uses a more intuitive approach, treating the observations as a random sample population, and allocating random effects accordingly.

1.12.1 Standard Two Way ANOVA

? 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 rth replicate measurement on the ith item by the mth method ($m = 1, 2$, $i = 1, \ldots, N$, and $r = 1, \ldots, n$), can be written as

$$y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + \epsilon_{mir}. \tag{1.31}$$

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 $\varepsilon \sim \mathcal{N}(0, \varphi_m^2)$. 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 (1.31) 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, **?** 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.

1.12.2 BXC2004 Model

? 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;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2) \tag{1.32}$$

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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \tag{1.33}$$

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.

1.12.3 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 mdoel that combined fixed and random effect terms formulated by **?** as follows;

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

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

1.13 Agreement Criteria

? 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 due to the disagreement in either between-item variabilities or within-item variabilities, or both. **?** allows for a formal test of each.

? 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 $\mathrm{H}_1 \colon \alpha_1 = \alpha_2$ and $\mathrm{H}_2 \colon \sigma_1^2 = \sigma_2^2$ and $\mathrm{H}_3 \colon g_1^2 = g_2^2$ hold simultaneously. **?** uses a Bonferroni correction to control the familywise error rate for tests of $\{\mathrm{H}_1, \mathrm{H}_2, \mathrm{H}_3\}$ and account for difficulties arising due to multiple testing. Roy also integrates H_2 and H_3 into a single testable hypothesis $\mathrm{H}_4 \colon \omega_1^2 = \omega_2^2$, where $\omega_m^2 = \sigma_m^2 + g_m^2$ represent the overall variability of method m. Disagreement in overall variability may be caused by different between-item variabilities, by different within-item variabilities, or by both. If the exact cause of disagreement between the two methods is not of interest, then the overall variability test H_4 is an alternative to testing H_2 and H_3 separately.

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

Model Terms (Roy 2009)

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

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

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

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

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

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

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

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

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

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

2.1 Model Formula

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

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

Here β_0 and β_m are fixed-effect terms representing, respectively, a model intercept and an overall effect for method m. The b_{1i} and b_{2i} terms represent random effect parameters corresponding to the two methods, having $\mathrm{E}(b_{mi}) = 0$ with $\mathrm{Var}(b_{mi}) = g_m^2$ and $\mathrm{Cov}(b_{mi}, b_{m'i}) = g_{12}$. The random error term for each response is denoted ϵ_{mir} having $\mathrm{E}(\epsilon_{mir}) = 0$, $\mathrm{Var}(\epsilon_{mir}) = \sigma_m^2$, $\mathrm{Cov}(b_{mir}, b_{m'ir}) = \sigma_{12}$, $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$ and $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{m'ir'}) = 0$. When two methods of measurement are in agreement, there is no significant differences between β_1 and β_2, g_1^2 andg_2^2, and σ_1^2 andσ_2^2. Here β_0 and β_m are

fixed-effect terms representing, respectively, a model intercept and an overall effect for

method m. The model can be reparameterized by gathering the β terms together into (fixed effect) intercept terms $\alpha_m = \beta_0 + \beta_m$. The b_{1i} and b_{2i} terms are correlated random effect parameters having $\mathrm{E}(b_{mi}) = 0$ with $\mathrm{Var}(b_{mi}) = g_m^2$ and $\mathrm{Cov}(b_{1i}, b_{2i}) = g_{12}$. The random error term for each response is denoted ϵ_{mir} having $\mathrm{E}(\epsilon_{mir}) = 0$, $\mathrm{Var}(\epsilon_{mir}) = \sigma_m^2$, $\mathrm{Cov}(\epsilon_{1ir}, \epsilon_{2ir}) = \sigma_{12}$, $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$ and $\mathrm{Cov}(\epsilon_{1ir}, \epsilon_{2ir'}) = 0$. Two methods of measurement are in complete agreement if the null hypotheses $\mathrm{H}_1 \colon \alpha_1 = \alpha_2$ and $\mathrm{H}_2 \colon \sigma_1^2 = \sigma_2^2$ and $\mathrm{H}_3 \colon g_1^2 = g_2^2$ hold simultaneously. **?** uses a Bonferroni correction to control the familywise error rate for tests of $\{\mathrm{H}_1, \mathrm{H}_2, \mathrm{H}_3\}$ and account for difficulties arising due to multiple testing. Roy also integrates H_2 and H_3 into a single testable hypothesis $\mathrm{H}_4 \colon \omega_1^2 = \omega_2^2$, where $\omega_m^2 = \sigma_m^2 + g_m^2$ represent the overall variability of method m. Disagreement in overall variability may be caused by different between-item variabilities, by different within-item variabilities, or by both. If the exact cause of disagreement between the two methods is not of interest, then the overall variability test H_4 is an alternative to testing H_2 and H_3 separately.

2.2 Formulation of the response vector

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

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

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

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

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

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

2.3 Decomposition of the response covariance matrix

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

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

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

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

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

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

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

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

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

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

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. **?** 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

? 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.

2.6 Repeated Measurements

In cases where there are repeated measurements by each of the two methods on the same subjects , Bland Altman suggest calculating the mean for each method on each

subject and use these pairs of means to compare the two methods.
The estimate of bias will be unaffected using this approach, but the estimate of the standard deviation of the differences will be too small, because of the reduction of the effect of repeated measurement error. Bland Altman propose a correction for this.
Carstensen attends to this issue also, adding that another approach would be to treat each repeated measurement separately.

In this model , the variances of the random effects must depend on m, since the different methods do not necessarily measure on the same scale, and different methods naturally must be assumed to have different variances. **?** attends to the issue of comparative variances.

2.7 LME

Consistent with the conventions of mixed models, (**?**) formulates the measurement $y_{ij}$from method i on individual j as follows;

$$y_{ij} = P_{ij}\theta + W_{ij}v_i + X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{2.4}$$

The design matrix P_{ij} , with its associated column vector θ, specifies the fixed effects common to both methods. The fixed effect specific to the jth method is articulated by the design matrix W_{ij} and its column vector v_i. The random effects common to both methods is specified in the design matrix X_{ij}, with vector b_j whereas the random effects specific to the ith subject by the jth method is expressed by Z_{ij}, and vector u_j. Noticeably this notation is not consistent with that described previously. The design matrices are specified so as to includes a fixed intercept for each method, and a random intercept for each individual. Additional assumptions must also be specified;

$$v_{ij} \sim N(0, \Sigma), \tag{2.5}$$

These vectors are assumed to be independent for different is, and are also mutually independent. All Covariance matrices are positive definite. In the above model effects

can be classed as those common to both methods, and those that vary with method. When considering differences, the effects common to both effectively cancel each other out. The differences of each pair of measurements can be specified as following;

$$d_{ij} = X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{2.6}$$

This formulation has seperate distributional assumption from the model stated previously.

This agreement covariate x is the key step in how this methodology assesses agreement.

2.8 Definition of Replicate measurements

Further to **?**, a formal definition is required of what exactly replicate measurements are

By replicates we mean two or more measurements on the same individual taken in identical conditions. In general this requirement means that the measurements are taken in quick succession.

? also remark that an important feature of replicate observations is that they should be independent of each other. This issue is addressed by **?**, in terms of exchangeability and linkage. Carstenen advises that repeated measurements come in two *substantially different* forms, depending on the circumstances of their measurement: exchangable and linked.

2.9 Exchangeable measurements

Repeated measurements are said to be exchangeable if no relationship exists between successive measurements across measurements. If the condition of exchangeability exists, a group of measurement of the same item determined by the same method can be re-arranged in any permutation without prejudice to proper analysis. There is no

reason to believe that the true value of the underlying variable has changed over the course of the measurements.

For the purposes of method comparison studies the following remarks can be made. The $r-$th measurement made by method 1 has no special correspondence to the $r-$th measurement made by method 2, and consequently any pairing of repeated measurements are as good as each other.

Exchangeable repeated measurements can be treated as true replicates.

2.10 Linked measurements

Repeated measurements are said to be linked if a direct correspondence exists between successive measurements across measurements, i.e. pairing. Such measurements are commonly made with a time interval between them, but simultaneously for both methods. Paired measurements are exchangeable, but individual measurements are not.

If the paired measurements are taken in a short period of time so that no real systemic changes can take place on each item, they can be considered true replicates. Should enough time elapse for systemic changes, linked repeated measurements can not be treated as true replicates.

2.11 Replicate measurements in Roy's paper

? takes its definition of replicate measurement: two or more measurements on the same item taken under identical conditions. Roy also assumes linked measurements, but it is can be used for the non-linked case.

2.12 Random effects

Further to **?**, if the measurements by a method on an item are not necessarily true replications, e.g., repeated measures over time, then additional terms may be needed for e_{mir}. **?** also addresses this issue by the addition of an interaction term (i.e. a random effect) $u_m i$, yielding

$$y_{mir} = \alpha_{mi} + u_{mi} + e_{mi}.$$

The additional interaction term is characterized as $u_{mi} \sim \mathcal{N}(0, \tau_m^2)$ (**?**).

This extra interaction term provides a source of extra variability, but this variance is not relevant to computing the case-wise differences.

? advises that the formulation of the model should take the exchangeability (in other words, whether or not the measurements are 'true replicates') into account. If there is a linkage between measurements (therefore not 'true' replicates) , the 'item by replicate' should be included in the model. If there is no linkage, and the replicates are indeed true replicates, the interaction term should be omitted.

? demonstrates how to compute the limits of agreement for two methods in the case of linked measurements. As a surplus source of variability is excluded from the computation, the limits of agreement are not unduly wide, which would have been the case if the measurements were treated as true replicates.

? also assigns a random effect u_{mi} for each response y_{mir}. Importantly Roy's model assumes linkage.

2.13 Model for replicate measurements

We generalize the single measurement model for the replicate measurement case, by additionally specifying replicate values. Let y_{mir} be the $r-$th replicate measurement for subject "i" made by method "m". Further to **?** fixed effect can be expressed with a single term α_{mi}, which incorporate the true value μ_i.

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

Combining fixed effects (**?**), we write,

$$y_{mir} = \alpha_{mi} + e_{mir}.$$

The following assumptions are required

- e_{mir} is independent of the fixed effects with mean $\mathrm{E}(e_{mir}) = 0$.
- Further to **?** between-item and within-item variances $\mathrm{Var}(\alpha_{mi}) = \sigma^2_{Bm}$ and $\mathrm{Var}(e_{mir}) = \sigma^2_{Wm}$
- In keeping with **?**, these variance shall be considered as part of the between-item variance covariance matrix $\boldsymbol{D}$ and the within-item variance covariance matrix $\boldsymbol{\Sigma}$ respectively, and will be denoted accordingly (i.e. d^2_m and σ^2_m).
- Additionally, the total variability of method "m", denoted ω^2_m is the sum of the within-item and between-item variabilities.

$$\omega^2_m = d^2_m + \sigma^2_m$$

Chapter 3

Introduction to Roy's Procedure

? 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 (**?**).

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

Differences in between-subject variabilities of the two methods arise when one method is yielding average response levels for individuals than are more variable than the average response levels for the same sample of individuals taken by the other method. Differences in within-subject variabilities of the two methods arise when one method is yielding responses for an individual than are more variable than the responses for this same individual taken by the other method. The two methods of measurement

can be considered to agree, and subsequently can be used interchangeably, if all three null hypotheses are true.

3.1 Introduction to Roy's methodology

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

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

? 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, **?** 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.

? 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 (**?**).

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

3.3 Criteria

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

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

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

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 **?** affords the opportunity to expand upon Carstensen's approach. sectionLikelihood and estimation

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

Chapter 4

Likelihood Ratio Tests

4.1 Likelihood

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

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

- Maximum likelihood (ML) estimation is a method of obtaining parameter estimates by optimizing the likelihood function. The likelihood function is constructed as a function of the parameters in the specified model.

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

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.

4.2 Likelihood Ratio Tests

The relationship between the respective models presented by **?** 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 (**?**). Conversely, **?** advises that testing hypotheses on fixed-effect parameters should be based on ML estimation, and that using REML would not be appropriate in this context.

The likelihood ratio test is the procedure used to compare the fit of two models. For each candidate model, the '-2 log likelihood' ($M2LL$) is computed. The test statistic for each of the three hypothesis tests is the difference of the $M2LL$ for each pair of models. If the $p-$value in each of the respective tests exceed as significance level chosen by the analyst, then the null model must be rejected.

$$-2 \ln \Lambda_d = [M2LL \text{ under } H_0 \text{ model}] - [M2LL \text{ under } H_A \text{ model}] \tag{4.1}$$

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

$$\nu = [\text{ LRT df under } H_0 \text{ model}] - [\text{ LRT df under } H_A \text{ model}] \tag{4.2}$$

The probability distribution of the test statistic can be approximated by a chi-

square distribution with (ν_1 - ν_2) degrees of freedom, where ν_1 and ν_2 are the degrees of freedom of models 1 and 2 respectively.

Likelihood ratio tests are very simple to implement in R, simply use the 'anova()' commands. Sample output will be given for each variability test.

4.3 Likelihood Ratio Tests

The relationship between the respective models presented by **?** 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 (**?**). Conversely, **?** advises that testing hypotheses on fixed-effect parameters should be based on ML estimation, and that using REML would not be appropriate in this context.

4.3.1 Likelihood Ratio Tests

The relationship between the respective models presented by **?** 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 (**?**, 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 (**?**). Conversely, **?** advises that testing hypotheses on fixed-effect parameters should be based on ML estimation, and that using REML would not be appropriate in this context.

4.4 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 (**?**).

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

4.4.1 Likelihood Ratio Tests

The relationship between the respective models presented by **?** 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 (**?**). Conversely, **?** advises that testing hypotheses on fixed-effect parameters should be based on ML estimation, and that using REML would not be appropriate in this context.

4.5 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.

4.5.1 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 (**?**, 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.

4.5.2 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 (**?**, pg.86).

One should be aware that these p-values may be conservative. That is, the reported p-value may be greater than the true p-value for the test and, in some cases, it may be much greater.(**?**, pg.87).

4.5.3 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 liklihood), but only if the fixed terms in the two models are invariant, and both models have been fit by REML. Otherwise, the argument: method="ML" must be employed (ML = maximum likelihood).

- Example of a likelihood ratio test used to compare two models:
  `>anova(modelA, modelB)`

- The output will contain a p-value, and this should be used in conjunction with the AIC scores to judge which model is preferred. Lower AIC scores are better.

- Generally, likelihood ratio tests should be used to evaluate the significance of terms on the random effects portion of two nested models, and should not be used to determine the significance of the fixed effects.

- A simple way to more reliably test for the significance of fixed effects in an LME model is to use conditional F-tests, as implemented with the simple "anova" function. Example:
  `>anova(modelA)`

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

4.5.4 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.

4.5.5 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]

4.5.6 Likelihood Ratio Tests

The relationship between the respective models presented by **?** 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 (**?**). Conversely, **?** advises that testing hypotheses on fixed-effect parameters should be based on ML estimation, and that using REML would not be appropriate in this context.

- LMEs
- Likelihood and 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).

4.5.7 Likelihood Ratio Tests

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

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

4.5.8 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 accounts 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.

4.5.9 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 liklihood), but only if the fixed terms in the two models are invariant, and both models have been fit by REML. Otherwise, the argument: method="ML" must be employed (ML = maximum likelihood).

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

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

4.5.10 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.

4.5.11 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.

4.6 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.

Chapter 5

Model Specification

5.1 Model Specification for Roy's Hypotheses Tests

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

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

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

It is assumed that $\boldsymbol{b}_i \sim N(0, \boldsymbol{G})$, $\boldsymbol{\epsilon}_i$ is a matrix of random errors distributed as $N(0, \boldsymbol{R}_i)$ and that the random effects and residuals are independent of each other.

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

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

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

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

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

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

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

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

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

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

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

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

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

$\boldsymbol{R}_i$ is the variance covariance matrix for the residuals, i.e. the within-item sources of variation between both methods. Computational analysis of linear mixed effects models allow for the explicit analysis of both $\boldsymbol{G}$ and $\boldsymbol{R_i}$. The above terms can be used to express the variance covariance matrix $\boldsymbol{\Omega}_i$ for the responses on item i ,

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

5.2 G Component

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

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

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

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

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

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

5.3 R Component

$\boldsymbol{R}_i$ is the variance covariance matrix for the residuals, i.e. the within-item sources of variation between both methods. The matrix of random errors $\boldsymbol{\epsilon}_i$ is distributed as $\mathcal{N}_2(0, \boldsymbol{R}_i)$.

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

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

where σ_1^2 and σ_2^2 are the within-subject variances of the respective methods, and σ_{12} is the within-item covariance between the two methods. The within-item variance covariance matrix $\boldsymbol{\Sigma}$ is assumed to be the same for all replications. Again it is important to note that no special assumptions are made about the structure of the matrix. Computational analysis of linear mixed effects models allow for the explicit analysis of both $\boldsymbol{G}$ and $\boldsymbol{R_i}$.

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

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

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

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

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

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

The computation of the limits of agreement require that the variance of the difference of measurements. This variance is easily computable from the estimate of the Block - $\mathbf{\Omega}_i$ matrix. Lack of agreement can arise if there is a disagreement in overall variabilities. This may be due to due to the disagreement in either between-item variabilities or within-item variabilities, or both. **?** allows for a formal test of each.

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

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

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

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

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

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

5.4 Hamlett

$\boldsymbol{R}_i$ is the variance covariance matrix for the residuals, i.e. the within-item sources of variation between both methods. Computational analysis of linear mixed effects models allow for the explicit analysis of both $\boldsymbol{G}$ and $\boldsymbol{R_i}$.

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

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

? shows that $\boldsymbol{R}_i$ can be expressed as $\boldsymbol{R}_i = \boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}$. The partial within-item variance?covariance matrix of two methods at any replicate is denoted $\boldsymbol{\Sigma}$, where σ_1^2 and σ_2^2 are the within-subject variances of the respective methods, and σ_{12} is the within-item covariance between the two methods. It is assumed that the within-item variance?covariance matrix $\boldsymbol{\Sigma}$ is the same for all replications. Again it is important to note that no special assumptions are made about the structure of the matrix.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \tag{5.1}$$

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

5.5 For Expository Purposes

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

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

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

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

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

5.6 Kroneckor

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

Both covariance matrices can be written as follows;

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

and

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

? shows that $\boldsymbol{R}_i$ can be expressed as $\boldsymbol{R}_i = \boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}$. The partial within-item variance?covariance matrix of two methods at any replicate is denoted $\boldsymbol{\Sigma}$, where σ_1^2 and σ_2^2 are the within-subject variances of the respective methods, and σ_{12} is the

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

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \tag{5.2}$$

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

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

5.7 Overall Variability

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

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

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

The computation of the limits of agreement require that the variance of the difference of measurements. This variance is easily computable from the estimate of the Block - $\boldsymbol{\Omega}_i$ matrix. Lack of agreement can arise if there is a disagreement in overall variabilities. This may be due to due to the disagreement in either between-item variabilities or within-item variabilities, or both. **?** allows for a formal test of each.

5.8 Off-Diagonal Components in Roy's Model

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

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

σ_x^2 and σ_y^2 describe the level of measurement error associated with each of the measurement methods for a given item. Attention must be given to the off-diagonal elements of the matrix.

It is intuitive to consider the measurement error of the two methods as independent of each other.

5.9 Formal Testing

A formal test can be performed to test the hypothesis that the off-diagonal terms are zero.

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

5.10 Using LMEs for Method Comparison

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

? remarks upon 'by-hand' approaches advocated in **?** discouragingly, describing them as tedious, unnecessary and 'outdated'. Due to the prevalence of modern statistical software, **?** advocates the adoption of computer based approaches to method comparison studies, allowing the use of LME models that would not have been feasible otherwise. Rather than using the 'by hand' methods, estimates for required parameters can be gotten directly from output code. Furthermore, using computer approaches removes constraints, such as the need for the design to be perfectly balanced. In part this is due to the increased profile of LME models, and furthermore the availability of capable software. Additionally LME based approaches may utilise the diagnostic and influence analysis techniques that have been developed in recent times.

Chapter 6

Roy's hypothesis tests

6.1 Hypothesis Testing

Variability tests proposed by **?** affords the opportunity to expand upon Carstensen's approach. **?** 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.

6.2 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, **?** presents a formal test for inter-method bias. With the null and alternative hypothesis denoted H_1 and K_1 respectively, this test is formulated as

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

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

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

and a formal test for the equality of within-item variances.

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

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

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

These tests are complemented by the ability to 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.

6.3 Roy's variability tests

For the purposes of method comparison, Roy presents a methodology utilising linear mixed effects model. The formulation contains a Kronecker product covariance structure in a doubly multivariate setup. This methodology provides for the formal testing of inter-method bias, between-subject variability and within-subject variability of two methods. By doubly multivariate set up, Roy means that the information on each patient or item is multivariate at two levels, the number of methods and number of replicated measurements. Further to **?**, it is assumed that the replicates are linked over time. However it is easy to modify to the unlinked case.

Roy sets out three conditions for two methods to be considered in agreement. Firstly that there be no significant bias. Second that there is no difference in the between-subject variabilities, and lastly that there is no significant difference in the within-subject variabilities. Should both the second and third conditions be fulfilled, then the overall variabilities of both methods would be equal. Roy additionally uses the

overall correlation coefficient to provide extra information about the comparison, with a minimum of 0.82 being required.

Roy proposes a series of three tests on the variance components of an LME model. For these tests, four candidate models are constructed. The difference in the models are specifically in how the the D and Λ matrices are constructed, using either an unstructured form or a compound symmetry form. The first model is compared against each of three other models successively.

6.4 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.

6.5 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.

6.6 Correlation coefficient

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

one can be overlooked. Indeed Roy demonstrates that placing undue importance to it can lead to incorrect conclusions. **?** remarks that PROC MIXED only gives overall correlation coefficients, but not their variances. Similarly variance are not determinable in `R` as yet either. Consequently it is not possible to carry out inferences based on all overall correlation coefficients.

6.7 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.

6.8 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 $\mathbf{V}$.

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

$$\mathrm{H}_0 : g_1^2 = g_2^2$$
$$\mathrm{H}_1 : g_1^2 \neq g_2^2$$

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

$$\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2$$
$$\mathrm{H}_1 : \sigma_1^2 \neq \sigma_2^2$$

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

$$\mathrm{H}_0 : \omega_1^2 = \omega_2^2$$
$$\mathrm{H}_1 : \omega_1^2 \neq \omega_2^2$$

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.

6.9 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.

6.10 Variability test 2

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

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

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

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

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

6.11 Variability test 3

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

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

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

6.12 Omnibus testing

$$\begin{pmatrix} \omega_e^2 & \omega^{en} \\ \omega_{en} & \omega_n^2 \end{pmatrix} = \begin{pmatrix} \psi_e^2 & \psi^{en} \\ \psi_{en} & \psi_n^2 \end{pmatrix} + \begin{pmatrix} \sigma_e^2 & \sigma^{en} \\ \sigma_{en} & \sigma_n^2 \end{pmatrix} \tag{6.1}$$
$$\begin{pmatrix} \omega_2^1 & 0 \\ 0 & \omega_2^2 \end{pmatrix} = \begin{pmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

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 - $\boldsymbol{\Omega}_i$ matrix. Lack of agreement can arise if there is a disagreement in overall variabilities. This may be due to due to the disagreement in either between-item variabilities or within-item variabilities, or both. **?** allows for a formal test of each.

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \tag{6.2}$$

$$\begin{pmatrix} \omega_e^2 & \omega^{en} \\ \omega_{en} & \omega_n^2 \end{pmatrix} = \begin{pmatrix} \psi_e^2 & \psi^{en} \\ \psi_{en} & \psi_n^2 \end{pmatrix} + \begin{pmatrix} \sigma_e^2 & \sigma^{en} \\ \sigma_{en} & \sigma_n^2 \end{pmatrix} \tag{6.3}$$

6.13 Formal testing for covariances

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

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

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

As with the tests for variability, this test is performed by comparing a pair of model fits corresponding to the null and alternative hypothesis. In addition to testing the overall covariance, similar tests can be formulated for both the component variabilities if necessary.

6.13.1 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-$\boldsymbol{\Omega}_i$, but equally applicable to the component variabilities $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$);

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

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

ric structure with that of a model fitted using the compound symmetric structure is equivalent to a test of the equality of variance.

6.13.2 Variance-Covariance Structures

Independence

As though analyzed using between subjects analysis.

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

Compound Symmetry

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

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

Unstructured

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

Autoregressive

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

6.14 VC structures

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

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

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

$$\mathbf{\Psi} = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix} \tag{6.4}$$

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

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

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

$$\mathbf{\Psi} = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix} \tag{6.5}$$

6.15 LRTs

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

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

Chapter 7

LOAS

7.1 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 (**?**) 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 **??**.

? 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. **?** computes the limits of agreement to the case with replicate measurements by using LME models.

Further to **?**, the computation of the limits of agreement follows from the inter-method 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 - $\mathbf{\Omega}_i$ matrix, i.e.

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

? demonstrate statistical flaws with two approaches proposed by **?** for the purpose

of calculating the variance of the inter-method bias when replicate measurements are available. Instead, **?** 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.

? 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. **?** 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.

7.2 00-Limits of Agreement in LME models

? computes the limits of agreement to the case with repeated measurements by using LME models.

? 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.

7.3 Carstensen et al

? 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 **?** to take proper cognizance of the replicate measurements. **?** demonstrate statistical flaws with two approaches proposed by **?** 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.

7.4 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.

7.5 Carstensen's Limits of Agreement

? 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.

? 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.

? 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 **?**, the computation of the limits of agreement follows from the inter-

method 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.

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

7.6 Carstensen's Model

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

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + \epsilon_{mir}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \tag{7.1}$$

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 $i = 1, 2 \ldots$

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

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

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 $\text{E}(c_{mi}) = 0$ 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 theiir 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* (**?**).

? 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 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.

7.7 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.

? 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 **?** to take proper cognizance of the replicate measurements. **?** demonstrate statistical flaws with two approaches proposed by **?** 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.

? 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 (1.31). 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 (1.31), when covariance between methods is present.

? 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.

7.8 Limits of Agreement in LME models

? 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. **?** 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 $\boldsymbol{D}$ and within-subject variability $\boldsymbol{\Lambda}$ can be presented in matrix form,

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

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

$$\text{var}(y_A - y_B) = 2d^2 + \lambda_A^2 + \lambda_B^2. \tag{7.2}$$

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

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

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_A^2 + \lambda_B^2 - 2(d_{AB} + \lambda_{AB}) \tag{7.3}$$

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

$$\bar{y_A} - \bar{y_B} \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \tag{7.4}$$

7.9 Limits of Agreement in LME models

? uses LME models to determine the limits of agreement. The limits of agreement (**?**) are ubiquitous in method comparison studies.

7.10 BXC2008 uses - Limits of Agreement in LME models

? 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. **?** 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 $\boldsymbol{D}$ and within-subject variability $\boldsymbol{\Lambda}$ can be presented in matrix form,

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

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

$$\text{var}(y_A - y_B) = 2d^2 + \lambda_A^2 + \lambda_B^2. \tag{7.5}$$

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

7.11 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.

7.12 Interaction Terms in Model

? 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 over-estimation 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

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

7.13 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 - $\mathbf{\Omega}_i$ matrix, i.e.

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

? 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 (1.31). 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 (1.31), when covariance between methods is present.

7.14 BXC - Model Terms

- Let y_{mir} be the response of method m on the ith subject at the $r-$th replicate.
- Let $\boldsymbol{y}_{ir}$ be the 2×1 vector of measurements corresponding to the $i-$th subject at the $r-$th replicate.
- Let $\boldsymbol{y}_i$ be the $R_i \times 1$ vector of measurements corresponding to the $i-$th subject, where R_i is number of replicate measurements taken on item i.
- Let $\alpha_m i$ be the fixed effect parameter for method for subject i.
- Formally Roy uses a separate fixed effect parameter to describe the true value μ_i, but later combines it with the other fixed effects when implementing the model.
- Let u_{1i} and u_{2i} be the random effects corresponding to methods for item i.

- $\boldsymbol{\epsilon}_i$ is a n_i-dimensional vector comprised of residual components. For the blood pressure data $n_i = 85$.

- $\boldsymbol{\beta}$ is the solutions of the means of the two methods. In the LME output, the bias ad corresponding t-value and p-values are presented. This is relevant to Roy's first test.

7.15 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 - $\boldsymbol{\Omega}_i$ 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.

7.16 LOAs with Roy

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

variance of the differences is

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_A^2 + \lambda_B^2 - 2(d_{AB} + \lambda_{AB}) \tag{7.6}$$

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

$$\bar{y_A} - \bar{y_B} \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \tag{7.7}$$

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.

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

7.17 Differences Between Models

? 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. **?** 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 (1.31) 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, **?** 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.

7.18 Differences

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

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_A^2 + \lambda_B^2 - 2(d_{AB} + \lambda_{AB}) \tag{7.8}$$

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

$$\bar{y_A} - \bar{y_B} \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \tag{7.9}$$

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.

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

7.19 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.

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

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

? uses examples from **?** to be able to compare both types of analysis.

? 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 $\hat{\Sigma}$.

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \tag{7.10}$$

7.20 Difference Variance further to Carstensen

? 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 $\tau_1 = \tau_2 = \tau$(**?**).*

Another important point is that there is no covariance terms, so further to **?** the variance covariance matrices for between-item and within-item variability are respectively.

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

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

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

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

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

7.21 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.

7.22 Carstensen's Mixed Models

? proposes linear mixed effects models for deriving conversion calculations similar to Deming's regression, and for estimating variance components for measurements by different methods. The following model (in the authors own notation) is formulated as follows, where y_{mir} is the rth replicate measurement on subject i with method m.

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{7.11}$$

The intercept term α and the $\beta_m \mu_i$ term follow from **?**, expressing constant and proportional bias respectively , in the presence of a real value μ_i. c_{mi} is a interaction term to account for replicate, and e_{mir} is the residual associated with each observation. Since variances are specific to each method, this model can be fitted separately for each method.

The above formulation doesn't require the data set to be balanced. However, it does require a sufficient large number of replicates and measurements to overcome the problem of identifiability. The import of which is that more than two methods of measurement may be required to carry out the analysis. There is also the assumptions that observations of measurements by particular methods are exchangeable within subjects. (Exchangeability means that future samples from a population behaves like earlier samples).

? uses the above formula to predict observations for a specific individual i by method m;

$$BLUP_{mir} = \hat{\alpha_m} + \hat{\beta_m} \mu_i + c_{mi} \tag{7.12}$$

. Under the assumption that the μs are the true item values, this would be sufficient to estimate parameters. When that assumption doesn't hold, regression techniques (known as updating techniques) can be used additionally to determine the estimates. The assumption of exchangeability can be unrealistic in certain situations. **?** provides

an amended formulation which includes an extra interaction term ($d_{mr} \sim N(0, \omega_m^2)$to account for this.

? sets out a methodology of computing the limits of agreement based upon variance component estimates derived using linear mixed effects models. Measures of repeatability, a characteristic of individual methods of measurements, are also derived using this method.

? also advocates the use of linear mixed models in the study of method comparisons. The model is constructed 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;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad (e_{mi} \sim N(0, \sigma_m^2)) \tag{7.13}$$

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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{7.14}$$

? proposes a methodology to calculate prediction intervals in the presence of replicate measurements, overcoming problems associated with Bland-Altman methodology in this regard. It is not possible to estimate the interaction variance components τ_1^2 and τ_2^2 separately. Therefore it must be assumed that they are equal. The variance of the difference can be estimated as follows:

$$var(y_{1j} - y_{2j}) \tag{7.15}$$

Bibliography