Contents

Chapter 1

Fitting MCS Models with R

1.1 LME models in method comparison studies

Linear mixed effects (LME) models can facilitate greater understanding of the potential causes of bias and differences in precision between two sets of measurement. Lai and Shiao (2005) views the uses of linear mixed effects models as an expansion on the Bland-Altman methodology, rather than as a replacement. Carstensen et al. (2008) remarks that modern statistical computation, such as that used for LME models, greatly improve the efficiency of calculation compared to previous 'by-hand' methods. In this chapter various LME approaches to method comparison studies shall be examined.

Roy provides three case studies, using data sets well known in method comparison studies, to demonstrate how the methodology should be used.

The first case study is the Systolic blood pressure data, taken from Bland and Altman (1999).

1.2 (good)Implementation in R

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

variable 'BP' (i.e. blood pressure).

The third argument contain the random effects component of the formulation, describing the random effects, and their grouping structure. The `nlme` package provides a set of positive-definite matrices , the `pdMat` class, that can be used to specify a structure for the between-subject variance-covariance matrix for the random effects. For Roy's methodology, we will use the `pdSymm` and `pdCompSymm` to specify a symmetric structure and a compound symmetry structure respectively. A full discussion of these structures can be found in Pinheiro and Bates (1994, pg. 158).

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

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

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

For the blood pressure data used in Roy (2009a), all four candidate models are implemented by slight variations of this piece of code, specifying either `pdSymm` or `pdCompSymm` in the second line, and either `corSymm` or `corCompSymm` in the fourth line. For example, the second candidate model 'MCS2' is implemented with the same code as MCS1, except for the term `pdCompSymm` in the second line, rather than `pdSymm`.

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

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

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

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

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

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

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

1.3 Important Consideration for MCS

The key issue is that `nlme` allows for the particular specification of ARoy2009's Model, specifically direct specification of the VC matrices for within subject and between subject residuals.

The `lme4` package does not allow for ARoy2009's Model, for reasons that will identified shortly. To advance the ideas that eminate from ARoy2009s' paper, one is required to use the `nlme` context. However, to take advantage of the infrastructure already provided for `lme4` models, one may change the research question away from that of ARoy2009's paper. To this end, an exploration of what textbfinfluence.ME can accomplished is merited.

1.4 Demonstration of Roy's testing (Good)

Roy (2009b) provides three case studies, using data sets well known in method comparison studies, to demonstrate how the methodology should be used. The first two examples used are from the 'blood pressure' data set introduced by Bland and Altman (1999). The data set is a tabulation of simultaneous measurements of systolic blood pressure were made by each of two experienced observers (denoted 'J' and 'R') using a sphygmomanometer and by a semi-automatic blood pressure monitor (denoted 'S'). Three sets of readings were made in quick succession. Roy compares the 'J' and 'S' methods in the first of her examples.

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

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

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

A likelihood ratio test is perform to compare both candidate models. The log-likelihood of the null model is -2030.7, and for the alternative model -2030.8. The test statistic, presented with greater

precision than the log-likelihoods, is 0.1592. The $p-$value is 0.6958. Consequently we fail to reject the null model, and by extension, conclude that the hypothesis that methods J and S have the same between-subject variability. Thus the second of the criteria is fulfilled.

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

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

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

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

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

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

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

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

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

$$\hat{\boldsymbol{r}}_{\boldsymbol{\Omega}ii} = \begin{pmatrix} 1 & 0.7959 \\ 0.7959 & 1 \end{pmatrix}$$

$$\hat{\boldsymbol{r}}_{\boldsymbol{\Omega}ii} = \begin{pmatrix} 1 & 0.7959 \\ 0.7959 & 1 \end{pmatrix} \tag{1.1}$$

The off-diagonal blocks of the overall correlation matrix $\hat{\boldsymbol{r}}_{\boldsymbol{\Omega}ii'}$ present the correlation coefficients further to Hamlett et al. (2004).

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

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

Using Roy's Test to Identify cause of Lack of agreement

Barnhart specifies three conditions for method of measurement that are required for two methods of measurement to be considered in agreement.

(i) No Significant Inter-method bias

(ii) No significant Difference in Within-Subject Variance

(iii) No significant Difference in Within-Subject Variance

Roy (2009b) demonstrates a LME model specification, and a series of tests that look at each of these agreement criteria individually. If two methods of measuement lack agreement, the specific reason or reasons for this lack of agreement can be identified.

Roy (2009b) 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),

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

1.4.1 Matrix structures

Before discussing the tests, it is useful to point out the difference between symmetric form and compound symmetry form. Consider a generic matrix A,

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}. \tag{1.2}$$

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

1.5 Computation of LMEs using R

Pinheiro and Bates (1994) advises how to implement LME models in statistical software (ostensibly for S and S PLUS, but R is very similar). When tackling linear mixed effects models using the R language, a statistician can call upon the *lme* command found in the *nlme* package.This command fits a LME model to the data set using either Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML). ML may be referred to as 'full maximum likelihood' estimation.

The first two arguments for *lme* are *fixed* and *data*, which give the model for the expected responses (i.e. the fixed part of the model), and the data that themodel should be fitted from. The next argument

is *random*, a one-sided formula which describes the random effects, and the grouping structure for the model. The *method* argument can specify whether to use 'REML', the default setting, or 'ML'.

1.6 Introduction

Roy (2009b) uses an approach based on linear mixed effects (LME) models for the purpose of comparing the agreement between two methods of measurement, where replicate measurements on items, typically 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.

1.7 Model terms

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

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

- Later on $\boldsymbol{X}_i$ will be reduced to a 2×1 matrix, to allow estimation of terms. This is due to a

shortage of rank. The fixed effects vector can be modified accordingly.

- $\boldsymbol{Z}_i$ is the $2n_i \times 2$ model matrix for the random effects for measurement methods on item i.
- $\boldsymbol{b}_i$ is the 2×1 vector of random-effect coefficients on item i, one for each method.
- $\boldsymbol{\epsilon}$ is the $2n_i \times 1$ vector of residuals for measurements on item i.
- $\boldsymbol{G}$ is the 2×2 covariance matrix for the random effects.
- $\boldsymbol{R}_i$ is the $2n_i \times 2n_i$ covariance matrix for the residuals on item i.
- The expected value is given as $\mathrm{E}(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$. (Hamlett et al., 2004)
- The variance of the response vector is given by $\mathrm{Var}(\boldsymbol{y}_i) = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R}_i$ (Hamlett et al., 2004).

1.8 Fitting Models with the LME4 R package

Two LME models are fitted to the data, one using the nlme package, one with the lme4 package. These models shall be called "blood.nlme" and "blood.lme4" respectively. In both cases the method is characterized by a fixed effect, while there is a random effect for each subject. This random effect accounts for the replicate measurements associated with each subject. The differences between the estimate provided by the respective models are negligible, due to the simplicity of the model specification.

Maximum likelihood or restricted maximum likelihood (REML) estimates of the parameters in linear mixed-effects models can be determined using the `lmer` function in the lme4 package for R. As for most model-fitting functions in R, the model is described in an `lmer` call by a formula, in this case including both fixed- and random-effects terms.

The formula and data together determine a numerical representation of the model from which the profiled deviance or the profiled REML criterion can be evaluated as a function of some of the model parameters. The appropriate criterion is optimized, using one of the constrained optimization functions in `R`, to provide the parameter estimates. We describe the structure of the model, the steps in evaluating the profiled deviance or REML criterion, and the structure of classes or types that represents such a model.

Sufficient detail is included to allow specialization of these structures by users who wish to write functions to fit specialized linear mixed models, such as models incorporating pedigrees or smoothing splines, that are not easily expressible in the formula language used by lmer.

`y` : Response variable

`method` : Method of Measurement

`subject` : Subject

`MCSdata`

```
library(lme4)

MCS.lme4 <- lmer(y ~ method-1 + (1|subject) , data=MCSdata)
```

1.8.1 Demonstration of Roy's testing

Roy provides three case studies, using data sets well known in method comparison studies, to demonstrate how the methodology should be used. The first two examples used are from the 'blood pressure' data set introduced by Bland and Altman (1999). The data set is a tabulation of simultaneous measurements of systolic blood pressure were made by each of two experienced observers (denoted 'J' and 'R') using a sphygmomanometer and by a semi-automatic blood pressure monitor (denoted 'S'). Three sets of readings were made in quick succession. Roy compares the 'J' and 'S' methods in the first of her examples.

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

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

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

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

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

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

Again, A likelihood ratio test is perform to compare both candidate models. The log-likelihood of the alternative model model is -2045.0. As before, the null model has a log-likelihood of -2030.7. The

test statistic is computed as 28.617, again presented with greater precision. The $p-$value is less than 0.0001. In this case we reject the null hypothesis of equal within-subject variability. Consequently the third of Roy's criteria is unfulfilled. The coefficient of repeatability for methods J and S are found to be 16.95 mmHg and 25.28 mmHg respectively.

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

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

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

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

$$\hat{\boldsymbol{r}}_{\boldsymbol{\Omega}ii} = \begin{pmatrix} 1 & 0.7959 \\ 0.7959 & 1 \end{pmatrix}$$

The off-diagonal blocks of the overall correlation matrix $\hat{\boldsymbol{r}}_{\boldsymbol{\Omega}ii'}$ present the correlation coefficients further to Hamlett et al. (2004).

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

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

1.9 Implementation in R

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

The third argument contain the random effects component of the formulation, describing the random effects, and their grouping structure. The `nlme` package provides a set of positive-definite matrices , the `pdMat` class, that can be used to specify a structure for the between-subject variance-covariance matrix for the random effects. For ARoy2009's methodology, we will use the `pdSymm` and `pdCompSymm` to specify a symmetric structure and a compound symmetry structure respectively. A full discussion of these structures can be found in Pinheiro and Bates (1994, pg. 158).

Similarly a variety of structures for the with-subject variance-covariance matrix can be implemented using `nlme`. To implement a particular matrix structure, one must specify both a variance function and correlation structure accordingly. Variance functions are used to model the variance structure of the within-subject errors. `varIdent` is a variance function object used to allow different variances according to the levels of a classification factor in the data. A compound symmetry structure is implemented using the `corCompSymm` class, while the symmetric form is specified by `corSymm` class. Finally, the estimation methods is specified as "ML" or "REML". The first of ARoy2009's candidate model can be implemented using the following code;

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

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

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

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

To perform a likelihood ratio test for two candidate models, simply use the `anova` command with the names of the candidate models as arguments. The following piece of code implement the first of ARoy2009's variability tests.

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

The fixed effects estimates are the same for all four candidate models. The inter-method bias can be easily determined by inspecting a summary of any model. The summary presents estimates for all of the important parameters, but not the complete variance-covariance matrices (although some simple

R functions can be written to overcome this). The variance estimates for the random effects for MCS2 is presented below.

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

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

1.9.1 Computation

Modern software packages can be used to fit models accordingly. The best linear unbiased predictor (BLUP) for a specific subject i measured with method m has the form $BLUP_{mir} = \hat{\alpha_m} + \hat{\beta_m}\mu_i + c_{mi}$, under the assumption that the μs are the true item values.

1.10 Roy's Methodology

For the purposes of comparing two methods of measurement, Roy (2009b) presents a methodology utilizing linear mixed effects model. This methodology provides for the formal testing of inter-method bias, between-subject variability and within-subject variability of two methods. The formulation contains a Kronecker product covariance structure in a doubly multivariate setup. By doubly multivariate set up, Roy means that the information on each patient or item is multivariate at two levels, the number of

methods and number of replicated measurements. Further to Lam et al. (1999), it is assumed that the replicates are linked over time. However it is easy to modify to the unlinked case.

Roy (2009b) sets out three criteria for two methods to be considered in agreement. Firstly that there be no significant bias. Second that there is no difference in the between-subject variabilities, and lastly that there is no significant difference in the within-subject variabilities. ARoy2009 further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal.

A formal test for inter-method bias can be implemented by examining the fixed effects of the model. This is common to well known classical linear model methodologies. The null hypotheses, that both methods have the same mean, which is tested against the alternative hypothesis, that both methods have different means. The inter-method bias and necessary $t-$value and $p-$value are presented in computer output. A decision on whether the first of ARoy2009's criteria is fulfilled can be based on these values.

Importantly Roy (2009b) further proposes a series of three tests on the variance components of an LME model, which allow decisions on the second and third of ARoy2009's criteria. For these tests, four candidate LME models are constructed. The differences in the models are specifically in how the the D and Λ matrices are constructed, using either an unstructured form or a compound symmetry form. To illustrate these differences, consider a generic matrix A,

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

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

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

1.11 Roy's LME methodology for assessing agreement

Barnhart et al. (2007) describes the sources of disagreement as differing population means, different between-subject variances, different within-subject variances between two methods and poor correlation between measurements of two methods.

Roy (2009b) proposes a novel method using the LME model with Kronecker product covariance structure in a doubly multivariate set-up to assess the agreement between a new method and an established method with unbalanced data and with unequal replications for different subjects,proposes the use of LME models to perform a test on two methods of agreement to determine whether they can be used interchangeably.

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 $\mathbf{V}$.

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

1.11.1 Roy's variability tests

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

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

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.

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

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

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

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \tag{1.3}$$

Roy (2009b) considers four independent hypothesis tests.

- 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.

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

1.12 Roy's Hypotheses Tests

In order to express ARoy2009'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.

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.

$\boldsymbol{G}$ 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:

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

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 $\boldsymbol{R_i}$.

Hamlett et al. (2004) shows that $\boldsymbol{R}_i$ can be expressed as $\boldsymbol{R}_i = \boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}$. The partial within-item variance?covariance matrix of two methods at any replicate is denoted $\boldsymbol{\Sigma}$, where σ_1^2 and σ_2^2 are the within-subject variances of the respective methods, and σ_{12} is the within-item covariance between the two methods. It is assumed that the within-item variance?covariance matrix $\boldsymbol{\Sigma}$ is the same for all replications. Again it is important to note that no special assumptions are made about the structure of the matrix.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \tag{1.4}$$

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

$$\boldsymbol{y}_i = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} + \begin{pmatrix} \epsilon_{1i1} \\ \epsilon_{2i1} \\ \epsilon_{1i2} \\ \epsilon_{2i2} \\ \epsilon_{1i3} \\ \epsilon_{2i3} \end{pmatrix}, \tag{1.5}$$

where

$$\boldsymbol{G} =$$

and

$$\boldsymbol{R}_i =$$

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

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

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

1.13 Model Terms for Roy's Techniques

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

$$\boldsymbol{b}_i = \begin{pmatrix} b_{1i} \\ b_{21} \end{pmatrix} \tag{1.6}$$

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

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

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

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

1.14 Basic Models Fits

Further to Pinheiro and Bates (1994), several simple LME models are constructed for the blood pressure data. This data set is the subject of a method comparison study in Bland and Altman (1999).

1.14.1 Implementing the Mixed Models Fits

They are implemented using the following `R` code, utilising the 'nlme' package. An analysis of variance is used to compare the model fits.

The `R` script:

```
fit1 = lme( BP ~ method, data = dat, random = ~1 | subject )
fit2 = update(fit1, random = ~1 | subject/method )
fit3 = update(fit1, random = ~method - 1 | subject )
#analysis of variance
anova(fit1,fit2,fit3)
```

1. Simplest workable model, allows differences between methods and incorporates a random intercept for each subject. For subject 1 we have

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{Z}_i = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \boldsymbol{b}_i = b$$

   where $\mathrm{E}(b) = 0$ and $\mathrm{var}(b) = \psi$.

2.

$$\boldsymbol{Z}_i = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \quad \boldsymbol{b}_i = \begin{pmatrix} b_1 & 0 \\ 0 & b_2 \end{pmatrix}$$

   where $\mathrm{E}(b_i) = 0$ and $\mathrm{var}(\boldsymbol{b}) = \boldsymbol{\Psi}$.

   The variance of error terms is a 6×6 matrix.

1.14.2 Roy's Reference Model

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

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

1.14.3 Model Fit 1

This is a simple model with no interactions. There is a fixed effect for each method and a random effect for each subject.

$$y_{ijk} = \beta_j + b_i + \epsilon_{ijk}, \qquad i = 1, \ldots, 2, j = 1, \ldots, 85, k = 1, \ldots, 3$$

$$b_i \sim \mathcal{N}(0, \sigma_b^2), \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

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

```
Linear mixed-effects model fit by REML
Data: dat
Log-restricted-likelihood: -2155.853
Fixed: BP ~ method
(Intercept)     methodS
127.40784    15.61961

Random effects:
Formula: ~1 | subject
(Intercept) Residual
StdDev:    29.39085 12.44454

Number of Observations: 510
Number of Groups: 85
```

The following output was obtained.

```
Linear mixed-effects model fit by REML
Data: dat
Log-restricted-likelihood: -2047.582
Fixed: BP ~ method
(Intercept)     methodS
127.40784    15.61961

Random effects:
Formula: ~method - 1 | subject
Structure: General positive-definite, Log-Cholesky parametrization
StdDev    Corr
methodJ  30.455093 methdJ
methodS  31.477237 0.835
Residual  7.763666

Number of Observations: 510
Number of Groups: 85


Linear mixed-effects model fit by REML
Data: dat
Log-restricted-likelihood: -2047.714
Fixed: BP ~ method
(Intercept)     methodS
127.40784    15.61961

Random effects:
Formula: ~1 | subject
```

```
(Intercept)
StdDev:    28.28452

Formula: ~1 | method %in% subject
(Intercept) Residual
StdDev:    12.61562 7.763666

Number of Observations: 510
Number of Groups:
subject method %in% subject
85                 170
```

Nested Model (Between-Item Variability)

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

1.14.4 Variability test 1

This is a test on whether both methods A and B have the same between-subject variability or not.

$$H_0 : \ d_A = d_B \tag{1.7}$$

$$H_A : \ d_A \neq d_B \tag{1.8}$$

When implemented using R, 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.

Bland-Altman's blood data

With the alternative model, the MLE of the between-subject variance covariance matrix is given by

$$\hat{\boldsymbol{D}}_{\boldsymbol{Symm}} = \begin{pmatrix} 923.98 & 785.24 \\ 785.24 & 971.30 \end{pmatrix} \tag{1.9}$$

With the null model the MLE is as follows:

$$\hat{\boldsymbol{D}}_{\boldsymbol{CS}} = \begin{pmatrix} 946.50 & 784.32 \\ 784.32 & 946.50 \end{pmatrix} \tag{1.10}$$

A likelihood ratio test is perform to determine which model is more suitable. The outcome of this test is presented in the following `R` code.

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

The test statistic is the difference of the -2 log likelihoods; 0.15291. The $p-$value is 0.6958. Therefore we fail to reject the hypothesis that both have the same between-subject variabilities.

1.14.5 Model Fit 2

Nested Model (Between-Item Variability)

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

Nested Model (Within item Variability)

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

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

This is a simple model, this time with an interaction effect. There is a fixed effect for each method. This model has random effects at two levels b_i for the subject, and another, b_{ij}, for the respective method within each subject.

$$y_{ijk} = \beta_j + b_i + b_{ij} + \epsilon_{ijk}, \qquad i = 1, \ldots, 2, j = 1, \ldots, 85, k = 1, \ldots, 3$$

$$b_i \sim \mathcal{N}(0, \sigma_1^2), \qquad b_{ij} \sim \mathcal{N}(0, \sigma_2^2), \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

In this model, the random interaction terms all have the same variance σ_2^2. These terms are assumed to be independent of each other, even within the same subject.

```
Linear mixed-effects model fit by REML
Data: dat
```

```
Log-restricted-likelihood: -2047.714
Fixed: BP ~ method
(Intercept)     methodS
127.40784    15.61961

Random effects:
Formula: ~1 | subject
(Intercept)
StdDev:    28.28452

Formula: ~1 | method %in% subject
(Intercept) Residual
StdDev:    12.61562 7.763666

Number of Observations: 510
Number of Groups:
subject method %in% subject
85                  170
```

1.14.6 Model Fit 2

This is a simple model, this time with an interaction effect. There is a fixed effect for each method. This model has random effects at two levels b_i for the subject, and another, b_{ij}, for the respective method within each subject.

$$y_{ijk} = \beta_j + b_i + b_{ij} + \epsilon_{ijk}, \qquad i = 1, \ldots, 2, j = 1, \ldots, 85, k = 1, \ldots, 3$$

$$b_i \sim \mathcal{N}(0, \sigma_1^2), \qquad b_{ij} \sim \mathcal{N}(0, \sigma_2^2), \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

In this model, the random interaction terms all have the same variance σ_2^2. These terms are assumed to be independent of each other, even within the same subject.

1.14.7 Variability test 2

This is a test on whether both methods A and B have the same within-subject variability or not.

$$H_0 : \ \lambda_A = \lambda_B \tag{1.11}$$

$$H_A : \ \lambda_A = \lambda_B \tag{1.12}$$

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.

Bland-Altman's blood data

For the null model the MLE of the within-subject variance covariance matrix is given below.

$$\hat{\boldsymbol{\Lambda}}_{\boldsymbol{Symm}} = \begin{pmatrix} 37.40 & 16.06 \\ 16.06 & 83.14 \end{pmatrix} \tag{1.13}$$

With the alternative model the MLE is as follows:

$$\hat{\boldsymbol{\Lambda}}_{\boldsymbol{CS}} = \begin{pmatrix} 60.27 & 16.06 \\ 16.06 & 60.27 \end{pmatrix} \tag{1.14}$$

A likelihood ratio test is perform to determine which model is more suitable. The outcome of this test is that it can be assumed that they have equal The test statistic is the difference of the -2 log likelihoods; 28.617. The $p-$value is less than 0.0001. In this case we reject the null hypothesis that both models have the same within-subject variabilities.

1.14.8 Model Fit 3

This model is a more general model, compared to 'model fit 2'. This model treats the random interactions for each subject as a vector and allows the variance-covariance matrix for that vector to be estimated from the set of all positive-definite matrices. $\boldsymbol{y_i}$ is the entire response vector for the ith

subject. $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ are the fixed- and random-effects design matrices respectively.

$$\boldsymbol{y_i} = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i} + \boldsymbol{\epsilon_i}, \qquad i = 1, \ldots, 85$$

$$\boldsymbol{Z_i} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}), \qquad \boldsymbol{\epsilon_i} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma^2\Lambda})$$

For the first subject the response vector, $\boldsymbol{y_1}$, is: The fixed effects design matrix $\boldsymbol{X_i}$ is given by:

observation	BP	subject	method	replicate
1	100.00	1	J	1
86	106.00	1	J	2
171	107.00	1	J	3
511	122.00	1	S	1
596	128.00	1	S	2
681	124.00	1	S	3

(Intercept)	method S
1	0
1	0
1	0
1	1
1	1
1	1

The random effects design matrix $\boldsymbol{Z_i}$ is given by:

1.14.9 Model Fit 3

This model is a more general model, compared to 'model fit 2'. This model treats the random interactions for each subject as a vector and allows the variance-covariance matrix for that vector to be

method J	method S
1	0
1	0
1	0
0	1
0	1
0	1

estimated from the set of all positive-definite matrices. $\boldsymbol{y_i}$ is the entire response vector for the ith subject. $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ are the fixed- and random-effects design matrices respectively.

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{b_i} + \boldsymbol{\epsilon_i}, \qquad i = 1, \ldots, 85$$

$$\boldsymbol{Z_i} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}), \qquad \boldsymbol{\epsilon_i} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma^2\Lambda})$$

For the first subject the response vector, $\boldsymbol{y_1}$, is:

observation	BP	subject	method	replicate
1	100.00	1	J	1
86	106.00	1	J	2
171	107.00	1	J	3
511	122.00	1	S	1
596	128.00	1	S	2
681	124.00	1	S	3

The fixed effects design matrix $\boldsymbol{X_i}$ is given by:

(Intercept)	method S
1	0
1	0
1	0
1	1
1	1
1	1

The random effects design matrix $\boldsymbol{Z_i}$ is given by:

method J	method S
1	0
1	0
1	0
0	1
0	1
0	1

1.14.10 Variability test 3

This is a test on whether both methods A and B have the same overall variability or not.

$$H_0 : \ \sigma_A = \sigma_B \tag{1.15}$$

$$H_A : \ \sigma_A = \sigma_B \tag{1.16}$$

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.

Bland-Altman's blood data

With the null model the MLE of the within-subject variance covariance matrix is given below.

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Symm}} = \begin{pmatrix} 961.38 & 801.40 \\ 801.40 & 1054.43 \end{pmatrix} \tag{1.17}$$

With the alternative model the MLE is as follows:

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{CS}} = \begin{pmatrix} 1007.92 & 801.65 \\ 801.65 & 1007.92 \end{pmatrix} \tag{1.18}$$

Again a likelihood ratio test is used to determine the most suitable of the two candidate models. The test statistic is the difference of the -2 log likelihoods; 28.884. The $p-$value is less than 0.0001. We again reject the null hypothesis. Each model has a different overall variability, a foregone conclusion from the second variability test.

1.14.11 Test for inter-method bias

The inter-method bias between the two method is found to be 15.62 , with a $p-$value of

1.15 Likelihood Ratio Test

A general method for comparing nested models fit by maximum liklihood is the liklihood 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 liklihood).

Example of a liklihood ratio test used to compare two models:

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, liklihood 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:

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

1.15.1 LRTs with R

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

$$-??2ln\Lambda_d = \text{[M2LL under H0 model]} - \text{[M2LL under HA model]} \tag{1.19}$$

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

$$\nu_{=}\text{[LRT df under H0 model]} - \text{[LRT df under HA model]} \tag{1.20}$$

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

```
#ANOVAs
test1 = anova(fit1,fit2) # Between-Subject Variabilities
test2 = anova(fit1,fit3) # Within-Subject Variabilities
test3 = anova(fit1,fit4) # Overall Variabilities
```

The matter of how well two methods of measurement are said to be in agreement is a frequently posed question in statistical literature. A useful, and broadly consistent, set of definitions of what this agreement entail is put forth by Barnhart et al and Roy (2009). As pointed out by earlier contributors to the subject (commonly referred to as Method Comparison Studies) Shared with previous contributions (Bland and Altman, Carstensen) is the condition that there should no systematic tendency for one of the methods to consistently provide a value higher that than of the other method. If such a tendency did exist, we would refer to it as an inter-method bias. In earlier literature, the emphasis was placed up on single measurements simultaneously by each of the methods of measurement. Several different approaches, such as the Bland-Altman plot, and Orthogonal Regression (a special case of Deming Regression where the residual variances are assumed to be equal) have been proposed. Arguably, for the single replicate case, the established methodologies are sufficient for assessing agreement between two methods. In subsequent contributions, the matter of assessing agreement in the presence of replicate measurements was addressed. Some approaches extended already established approaches (Bland-Altam 1999). Other contributions were based on methodologies not seen previously in Method comparison Study Literature (for example, Carstensen et al 2008 and Roy 2009, using LME models). A review of recent literature demonstrates how useful and effective the use of LME models are.

1.16 Worked Eamples : LikelihoodRatio Tests

1.16.1 Nested Model (Overall Variability)

Additionally there is a third nested model, that can be used to test overall variability, substantively a a joint test for between-item and within-item variability. The motivation for including such a test in the suite is not clear, although it does circumvent the need for multiple comparison procedures in certain circumstances, hence providing a simplified procedure for non-statisticians.

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

1.16.2 ANOVAs for Original Fits

The likelihood Ratio test is very simple to implement in R. All that is required it to specify the reference model and the relevant nested mode as arguments to the command `anova()`.

The figure below displays the three tests described by Roy (2009).

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

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

1.16.3 Nested Model (Between-Item Variability)

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

- **Blood (JSR) data:**
- **PEFR Data:** ARoy20092009
- **Oximetry data:** BXC2004
- **Fat data:** BXC2004
- **Trig Gerber Data:** BXC2008
- **Nadler Hurley:**

- **Hamlett:**

1.17 results

Using Carstensen's method, the standard deviations of the casewise differences were computed as 20.43139089,20.26824078,2.260886283 respectively. Using ARoy2009's model, these deviations are estimated to be 20.32756749, 20.16326412, 2.252869282 respectively.

Similarly for the fat and ox data Carstensen computes the difference deviations as and 6.16867323 , whereas under ARoy2009's model they are estimated to be 6.139275202 and respectively.

However, using the PEFR and cardiac data, differences emerge.

Data	Carstensen	ARoy2009
Fat	0.1352	0.1373
Ox	6.1686	6.1392
Blood JS	20.4314	20.3275
Blood JR	2.26088	2.2528
Blood RS	20.2682	20.16326412
Hamlett	0.9031	0.8922

Using Roy's Model to Compute LoAs and CR

In this short section, a demonstration of how ARoy2009's technique can be used to compute two common MCS metrics: Limits of Agreement and the Coefficient of Repeatabilty. While Limits of Agreement are not used in the analysis proposed here, they are ubiquituous in literature, and a demonstration on how to compute them with the ARoy2009 Model would assist the adoption of this proposed method.

The coefficient of repeatability is encountered in Gage R & R analysis. *(A future exploration of how LME models can be used in that field would be of interest. This is something to include in the Conclusions Section).*

1.18 Using REML Fitting

Noticeably Roy (2009) uses ML estimation when specifying the LME models. No explanation is given, although plausibly it is due to the constraints of the computational environment being used. Both West et al (2010) and Pinheiro and Bates (2000) compare ML and REML estimation, describing what types of tests are appropriate for each. When variance components are being tested, REML estimation is in fact the correct approach.

```
   > fit1r = lme(y ~ meth-1, data = dat,   #Symm , Symm#
+     random = list(item=pdSymm(~ meth-1)),
+     weights=varIdent(form=~1|meth),
+     correlation = corSymm(form=~1 | item/repl),
+     method="REML")



> fit2r = lme(y ~ meth-1, data = dat,   #CS , Symm#
+     random = list(item=pdCompSymm(~ meth-1)),
+     correlation = corSymm(form=~1 | item/repl),
+     method="REML")







> fit3r = lme(y ~ meth-1, data = dat,   #Symm , CS#
+     random = list(item=pdSymm(~ meth-1)),
+     weights=varIdent(form=~1|meth),
+     correlation = corCompSymm(form=~1 | item/repl),
+     method="REML")
```

```
> fit4r = lme(y ~ meth-1, data = dat,   #CS , CS#
+     random = list(item=pdCompSymm(~ meth-1)),
+     correlation = corCompSymm(form=~1 | item/repl),
+     method="REML")
```

```
> test1r = anova(fit1r,fit2r)                    # Between-Subject Variabilities
> test2r = anova(fit1r,fit3r)                   # Within-Subject Variabilities
> test3r = anova(fit1r,fit4r)                   # Overall Variabilities
```

```
> fit1bias = lme(y ~ meth, data = dat,   #Symm , Symm#
+     random = list(item=pdSymm(~ meth-1)),
+     weights=varIdent(form=~1|meth),
+     correlation = corSymm(form=~1 | item/repl),
+     method="ML")
```

Comparison of ML and REML fits

```
Fit 1 (ML)

Dataset: Blood RS

Fixed : 127.3126 , 143.0275

AIC: 4075.594

Between Subject Variability

  Fit1r (REML)

Dataset: Blood RS

Fixed : 127.3126 , 143.0275

AIC: 4068.172

Between Subject Variability
```

MAY 2012 : Research Notes Roy (2009) proposes a suite of hypothesis tests for assessing the agreement of two methods of measurement, when replicate measurements are obtained for each item, using a LME approach. (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. Testing 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.

1.18.1 Roy's Candidate Models

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

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

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

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

```
Comparison of ML and REML fits
Fit 1 (ML)

Dataset: Blood RS

Fixed : 127.3126 , 143.0275

AIC: 4075.594

Between Subject Variability

Fit1r (REML)

Dataset: Blood RS

Fixed : 127.3126 , 143.0275
```

```
AIC: 4068.172

Between Subject Variability
```

```
#Roy's Candidate Models


fit1 = lme(y ~ meth-1, data = dat,   #Symm , Symm#
random = list(item=pdSymm(~ meth-1)),
weights=varIdent(form=~1|meth),
correlation = corSymm(form=~1 | item/repl),
method="ML")

fit2 = lme(y ~ meth-1, data = dat,   #CS , Symm#
random = list(item=pdCompSymm(~ meth-1)),
correlation = corSymm(form=~1 | item/repl),
method="ML")

fit3 = lme(y ~ meth-1, data = dat,   #Symm , CS#
random = list(item=pdSymm(~ meth-1)),
weights=varIdent(form=~1|meth),
```

```
correlation = corCompSymm(form=~1 | item/repl),
method="ML")

fit4 = lme(y ~ meth-1, data = dat,   #CS , CS#
random = list(item=pdCompSymm(~ meth-1)),
correlation = corCompSymm(form=~1 | item/repl),
method="ML")
```

1.19 Replicates

```
Linear mixed-effects model fit by REML
Data: dat
Log-restricted-likelihood: -2047.714
Fixed: BP ~ method
(Intercept)     methodS
127.40784    15.61961

Random effects:
Formula: ~1 | subject
(Intercept)
StdDev:    28.28452

Formula: ~1 | method %in% subject
(Intercept) Residual
StdDev:    12.61562 7.763666
```

```
Number of Observations: 510
Number of Groups:
subject method %in% subject
85                  170
```

1.20 Roy's examples

To complete the study, the relevant values are provided for the Rvs$S$ comparison also.

The second data set, a comparison of two peak expiratory flow rate measurements, is referenced by Bland and Altman (1986).

The last case study is also based on a data set from Bland and Altman (1999). It contains the measurements of left ventricular cardiac eject fraction, measured by impedance cartography and radionuclide ventriculography, on twelve patients. The number of replicated differs for each patient.

The bias is shown to be 0.7040, with a p-value of 0.0204. The MLEa of the between-method and within-method variance-covariance matrices of methods RV and IC are given by

$$\hat{D} = \begin{pmatrix} 1.6323 & 1.1427 \\ 1.1427 & 1.4498 \end{pmatrix}, \tag{1.21}$$

$$\hat{\Sigma} = \begin{pmatrix} 1.6323 & 1.1427 \\ 1.1427 & 1.4498 \end{pmatrix}. \tag{1.22}$$

Roy (2009a) notes that these are the same estimate for variance as given by Bland and Altman (1999).

The repeatability coefficients are determined to be 0.9080 for the RV method and 1.0293 for the IC method.

From the estimated $\boldsymbol{\Omega_i}$ correlation matrix, the overall correlation coefficient is 0.7100. The overall correllation coefficients between two methods RV and IC are 0.9384 and 0.9131 respectively.

Roy (2009a) concludes that is appropriate to switch between the two methods if needed.

?

Roy (2009a) recommends to not switch between the two method.

1.21 LME

Fitting model according to Roy

```
Linear mixed-effects model fit by REML
Data: BA99
AIC      BIC    logLik
4319.707 4336.629 -2155.853

Random effects:
Formula: ~1 | subj
(Intercept) Residual
StdDev:    29.39085 12.44454

Fixed effects: ob.js ~ method
Value Std.Error  DF  t-value p-value
(Intercept) 127.40784  3.281757 424 38.82306       0

methodS     15.61961   1.102107 424 14.17250       0

Correlation:
(Intr)
methodS -0.168

Standardized Within-Group Residuals:
Min          Q1         Med          Q3         Max
-3.61292639 -0.42538402 -0.02467651  0.40166235  4.84280044

Number of Observations: 510 Number of Groups: 85
```

1.22 IC/RV comparison

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

$$\hat{D} = \begin{bmatrix} 1.6323 & 1.1427 \\ 1.1427 & 1.4498 \end{bmatrix} \tag{1.23}$$

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

$$\hat{\Sigma} = \begin{bmatrix} 0.1072 & 0.0372 \\ 0.0372 & 0.1379 \end{bmatrix} \tag{1.24}$$

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

$$Block\Omega_i = \begin{bmatrix} 1.7396 & 1.1799 \\ 1.1799 & 1.5877 \end{bmatrix}. \tag{1.25}$$

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

1.23 PEFR and Cardiac

Two further data sets applied to both methodologies are the"Cardiac" and "PEFR" , which are both contained on Carstensen's MethComp package. This data is from Bland and Altman (1986): two measurements of peak expiratory flow rate (PEFR) are compared. One of these measurements uses a "Large" meter and the other a "Mini" meter.

Two measurements were made with a Wright peak flow meter and two with a mini Wright meter, in random order. All measurements were taken by the same observer, using the same two instruments. (These data were collected to demonstrate the statistical method and provide no evidence on the comparability of these two instruments.)

Bibliography

Barnhart, H., M. Haber, and L. Lin (2007). An overview of assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics 17*, 529–569.

Bland, J. and D. Altman (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet i*, 307–310.

Bland, J. and D. Altman (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research 8*(2), 135–160.

Carstensen, B., J. Simpson, and L. C. Gurrin (2008). Statistical models for assessing agreement in method comparison studies with replicate measurements. *The International Journal of Biostatistics 4*(1).

Hamlett, A., L. Ryan, and R. Wolfinger (2004). On the use of PROC MIXED to estimate correlation in the presence of repeated measures. *Proceedings of the Statistics and Data Analysis Section, SAS Users Group International 198-229*, 1–7.

Lai, D. and S.-Y. P. K. Shiao (2005). Comparing two clinical measurements: a linear mixed model approach. *Journal of Applied Statistics 32*(8), 855–860.

Lam, M., K. Webb, and D. O'Donnell (1999). Correlation between two variables in repeated measurements. *American Statistical Association, Proceedings of the Biometric Session*, 213–218.

Pinheiro, J. and D. Bates (1994). *Mixed Effects Models in S and S plus* (2nd ed.). Reading, Massachusetts: Springer.

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

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