

Print Ready Draft for Chaps 1 and 2

Kevin O'Brien

January 21, 2015

Contents

Chapter 1

Introduction

1.1 Introduction

The problem of assessing the agreement between two or more methods of measurement is ubiquitous in scientific research, and is commonly referred to as a ‘method comparison study’. Published examples of method comparison studies can be found in disciplines as diverse as pharmacology (?), anaesthesia (?), and cardiac imaging methods (?).

To illustrate the characteristics of a typical method comparison study consider the data in Table I (?). In each of twelve experimental trials, a single round of ammunition was fired from a 155mm gun and its velocity was measured simultaneously (and independently) by three chronographs devices, identified here by the labels ‘Fotobalk’, ‘Counter’ and ‘Terma’.

Round	Fotobalk [F]	Counter [C]	Terma [T]
1	793.8	794.6	793.2
2	793.1	793.9	793.3
3	792.4	793.2	792.6
4	794.0	794.0	793.8
5	791.4	792.2	791.6
6	792.4	793.1	791.6
7	791.7	792.4	791.6
8	792.3	792.8	792.4
9	789.6	790.2	788.5
10	794.4	795.0	794.7
11	790.9	791.6	791.3
12	793.5	793.8	793.5

Table 1.1: Velocity measurement from the three chronographs (Grubbs 1973).

An important aspect of these data is that all three methods of measurement are assumed to have an attendant measurement error, and the velocities reported in Table 1.1 can not be assumed to be ‘true values’ in any absolute sense.

A method of measurement should ideally be both accurate and precise. ? describes agreement as being a broader term that contains both of those qualities. An accurate measurement method will give results close to the unknown ‘true value’. The precision of a method is indicated by how tightly measurements obtained under identical conditions are distributed around their mean measurement value. A precise and accurate method will yield results consistently close to the true value. Of course a method may be accurate, but not precise, if the average of its measurements is close to the true

value, but those measurements are highly dispersed. Conversely a method that is not accurate may be quite precise, as it consistently indicates the same level of inaccuracy. The tendency of a method of measurement to consistently give results above or below the true value is a source of systematic bias. The smaller the systematic bias, the greater the accuracy of the method.

In the context of the agreement of two methods, there is also a tendency of one measurement method to consistently give results above or below the other method. Lack of agreement is a consequence of the existence of ‘inter-method bias’. For two methods to be considered in good agreement, the inter-method bias should be in the region of zero. A simple estimation of the inter-method bias can be calculated using the differences of the paired measurements. The data in Table 1.2 are a good example of possible inter-method bias; the ‘Fotobalk’ consistently recording smaller velocities than the ‘Counter’ method. Consequently one would conclude that there is lack of agreement between the two methods.

The absence of inter-method bias by itself is not sufficient to establish whether two measurement methods agree. The two methods must also have equivalent levels of precision. Should one method yield results considerably more variable than those of the other, they can not be considered to be in agreement. With this in mind a methodology is required that allows an analyst to estimate the inter-method bias, and to compare the precision of both methods of measurement.

Round	Fotobalk (F)	Counter (C)	F-C
1	793.8	794.6	-0.8
2	793.1	793.9	-0.8
3	792.4	793.2	-0.8
4	794.0	794.0	0.0
5	791.4	792.2	-0.8
6	792.4	793.1	-0.7
7	791.7	792.4	-0.7
8	792.3	792.8	-0.5
9	789.6	790.2	-0.6
10	794.4	795.0	-0.6
11	790.9	791.6	-0.7
12	793.5	793.8	-0.3

Table 1.2: Difference between Fotobalk and Counter measurements.

1.2 Bland-Altman methodology

The issue of whether two measurement methods comparable to the extent that they can be used interchangeably with sufficient accuracy is encountered frequently in scientific research. Historically comparison of two methods of measurement was carried out by use of paired sample t -test, correlation coefficients or simple linear regression. Simple linear regression is unsuitable for method comparison studies because of the required assumption that one variable is measured without error. In comparing two methods, both methods are assumed to have attendant random error.

Statisticians Martin Bland and Douglas Altman recognized the inadequacies of these analyses and articulated quite thoroughly the basis on which they are unsuitable for comparing two methods of measurement (?). Furthermore they proposed their simple methodology specifically constructed for method comparison studies. They acknowledge the opportunity to apply other valid, but complex, methodologies, but argue that a simple approach is preferable, especially when the results must be ‘explained to non-statisticians’.

Notwithstanding previous remarks about linear regression, the first step recommended, which the authors argue should be mandatory, is construction of a simple scatter plot of the data. The line of equality should also be shown, as it is necessary to give the correct interpretation of how both methods compare. In the case of good agreement, the observations would be distributed closely along the line of equality. A scatter plot of the Grubbs data is shown in Figure 1.1. Visual inspection confirms the previous conclusion that there is an inter-method bias present, i.e. Fotobalk device has a tendency to record a lower velocity.

? notes that scatter plots were very seldom presented in the *Annals of Clinical Biochemistry*. This apparently results from the fact that the ‘Instructions for Authors’



Figure 1.1: Scatter plot For Fotobalk and Counter Methods.

dissuade the use of regression analysis, which conventionally is accompanied by a scatter plot.

1.2.1 Bland-Altman plots

In light of shortcomings associated with scatterplots, ? recommend a further analysis of the data. Firstly case-wise differences of measurements of two methods $d_i = y_{1i} - y_{2i}$ for $i = 1, 2, \dots, n$ on the same subject should be calculated, and then the average of those measurements ($a_i = (y_{1i} + y_{2i})/2$ for $i = 1, 2, \dots, n$).

? proposes a scatterplot of the case-wise averages and differences of two methods of measurement. This scatterplot has since become widely known as the Bland-Altman plot. ? express the motivation for this plot thusly:

“From this type of plot it is much easier to assess the magnitude of disagreement (both error and bias), spot outliers, and see whether there is any trend, for example an increase in (difference) for high values. This way of plotting the data is a very powerful way of displaying the results of a method comparison study.”

The case wise-averages capture several aspects of the data, such as expressing the range over which the values were taken, and assessing whether the assumptions of constant variance holds. Case-wise averages also allow the case-wise differences to be presented on a two-dimensional plot, with better data visualization qualities than a one dimensional plot. ? cautions that it would be the difference against either measurement value instead of their average, as the difference relates to both value. This methodology has proved very popular, and the Bland-Altman plots is widely regarded as powerful graphical methodology for making a visual assessment of the data.

The magnitude of the inter-method bias between the two methods is simply the average of the differences \bar{d} . This inter-method bias is represented with a line on the Bland-Altman plot. As the objective of the Bland-Altman plot is to advise on the agreement of two methods, it is the case-wise differences that are also particularly

relevant. The variances around this bias is estimated by the standard deviation of these differences S_d .

1.2.2 Bland-Altman plots for the Grubbs data

In the case of the Grubbs data the inter-method bias is -0.61 metres per second, and is indicated by the dashed line on Figure 1.2. By inspection of the plot, it is also possible to compare the precision of each method. Noticeably the differences tend to increase as the averages increase.

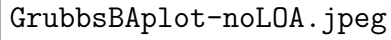
The Bland-Altman plot for comparing the ‘Fotobalk’ and ‘Counter’ methods, which shall henceforth be referred to as the ‘F vs C’ comparison, is depicted in Figure 1.2, using data from Table 1.3. The presence and magnitude of the inter-method bias is indicated by the dashed line.

Round	Fotobalk [F]	Counter [C]	Differences [F-C]	Averages [(F+C)/2]
1	793.8	794.6	-0.8	794.2
2	793.1	793.9	-0.8	793.5
3	792.4	793.2	-0.8	792.8
4	794.0	794.0	0.0	794.0
5	791.4	792.2	-0.8	791.8
6	792.4	793.1	-0.7	792.8
7	791.7	792.4	-0.7	792.0
8	792.3	792.8	-0.5	792.5
9	789.6	790.2	-0.6	789.9
10	794.4	795.0	-0.6	794.7
11	790.9	791.6	-0.7	791.2
12	793.5	793.8	-0.3	793.6

Table 1.3: Fotobalk and Counter methods: differences and averages.

Round	Fotobalk [F]	Terma [T]	Differences [F-T]	Averages [(F+T)/2]
1	793.8	793.2	0.6	793.5
2	793.1	793.3	-0.2	793.2
3	792.4	792.6	-0.2	792.5
4	794.0	793.8	0.2	793.9
5	791.4	791.6	-0.2	791.5
6	792.4	791.6	0.8	792.0
7	791.7	791.6	0.1	791.6
8	792.3	792.4	-0.1	792.3
9	789.6	788.5	1.1	789.0
10	794.4	794.7	-0.3	794.5
11	790.9	791.3	-0.4	791.1
12	793.5	793.5	0.0	793.5

Table 1.4: Fotobalk and Terma methods: differences and averages.



GrubbsBAplot-noLOA.jpeg

Figure 1.2: Bland-Altman plot For Fotobalk and Counter methods.

In Figure 1.3 Bland-Altman plots for the ‘F vs C’ and ‘F vs T’ comparisons are shown, where ‘F vs T’ refers to the comparison of the ‘Fotobalk’ and ‘Terma’ methods. Usage of the Bland-Altman plot can be demonstrate in the contrast between these comparisons. By inspection, there exists a larger inter-method bias in the ‘F vs C’ comparison than in the ‘F vs T’ comparison. Conversely there appears to be less precision in ‘F vs T’ comparison, as indicated by the greater dispersion of covariates.



Figure 1.3: Bland-Altman plots for Grubbs' F vs C and F vs T comparisons.

1.2.3 Prevalence of the Bland-Altman plot

?, which further develops the Bland-Altman methodology, was found to be the sixth most cited paper of all time by the ?. ? describes the rate at which prevalence of the Bland-Altman plot has developed in scientific literature. ? reviewed the use of Bland-Altman plots by examining all articles in the journal 'Clinical Chemistry' between 1995 and 2001. This study concluded that use of the BlandAltman plot increased over the years, from 8% in 1995 to 14% in 1996, and 3136% in 2002.

The Bland-Altman Plot has since become expected, and often obligatory, approach for presenting method comparison studies in many scientific journals (?). Furthermore ? recommend its use in papers pertaining to method comparison studies for the journal of the British Hypertension Society.

1.2.4 Adverse features

Estimates for inter-method bias and variance of differences are only meaningful if there is uniform inter-bias and variability throughout the range of measurements. Fulfilment of these assumptions can be checked by visual inspection of the plot. The prototype Bland-Altman plots depicted in Figures 1.4, 1.5 and 1.6 are derived from simulated data, for the purpose of demonstrating how the plot would inform an analyst of features that would adversely affect use of the recommended methodology.

Figure 1.4 demonstrates how the Bland-Altman plot would indicate increasing variance of differences over the measurement range. Fitted regression lines, for both the upper and lower half of the plot, has been added to indicate the trend. Figure 1.5 is an example of cases where the inter-method bias changes over the measurement range. This is known as proportional bias, and is defined by ? as meaning that ‘one method gives values that are higher (or lower) than those from the other by an amount that is proportional to the level of the measured variable’. In both Figures 1.4 and 1.5, the assumptions necessary for further analysis using the limits of agreement are violated.

Application of regression techniques to the Bland-Altman plot, and subsequent formal testing for the constant variability of differences is informative. The data set may be divided into two subsets, containing the observations wherein the difference values are less than and greater than the inter-method bias respectively. For both of these fits, hypothesis tests for the respective slopes can be performed. While both tests can be considered separately, multiple comparison procedures, such as the Benjamini-Hochberg (?) test, should be also be used.

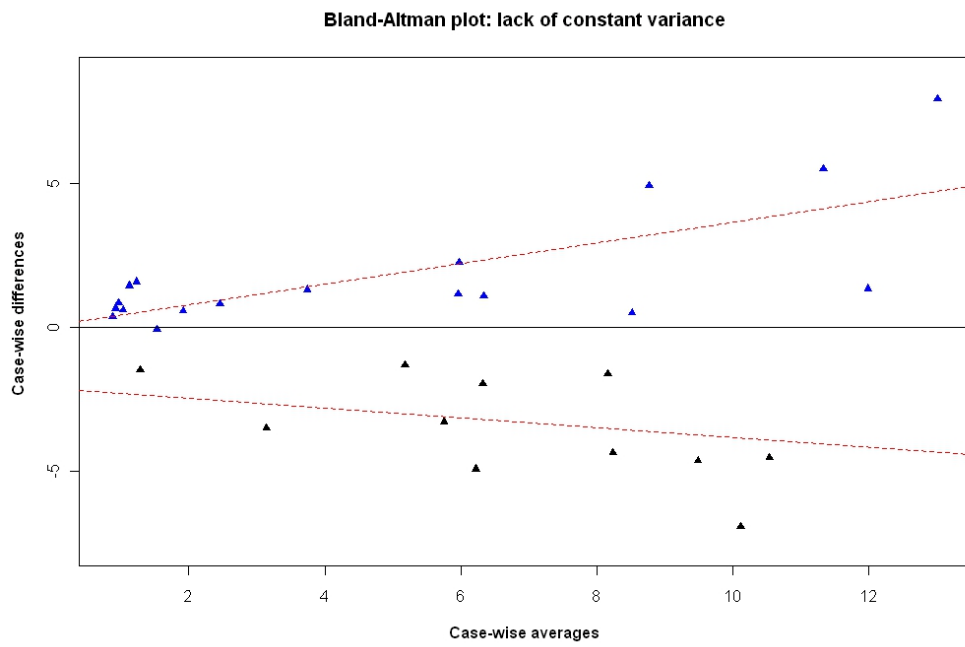


Figure 1.4: Bland-Altman plot demonstrating the increase of variance over the range.

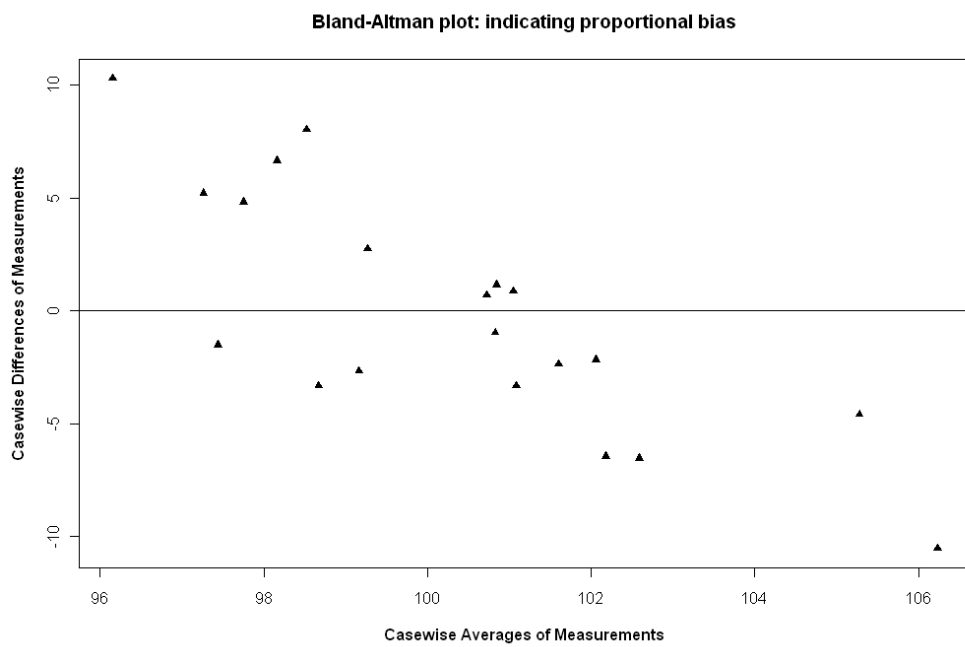


Figure 1.5: Bland-Altman plot indicating the presence of proportional bias.

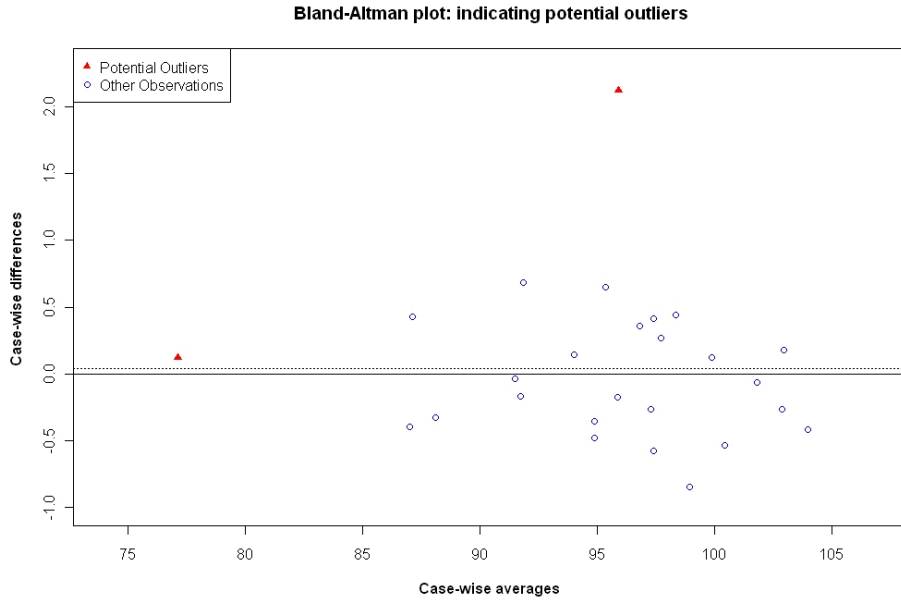


Figure 1.6: Bland-Altman plot indicating the presence of potential outliers.

The Bland-Altman plot also can be used to identify outliers. An outlier is an observation that is conspicuously different from the rest of the data that it arouses suspicion that it occurs due to a mechanism, or conditions, different to that of the rest of the observations. ? do not recommend excluding outliers from analyzes, but remark that recalculation of the inter-method bias estimate, and further calculations based upon that estimate, are useful for assessing the influence of outliers. The authors remark that ‘we usually find that this method of analysis is not too sensitive to one or two large outlying differences’. Figure 1.6 demonstrates how the Bland-Altman plot can be used to visually inspect the presence of potential outliers.

As a complement to the Bland-Altman plot, ? proposes the use of a bivariate confidence ellipse, constructed for a predetermined level. ? provides the relevant calculations for the ellipse. This ellipse is intended as a visual guidelines for the scatter plot, for detecting outliers and to assess the within- and between-subject variances.

The minor axis relates to the between subject variability, whereas the major axis

relates to the error mean square, with the ellipse depicting the size of both relative to each other. Consequently Bartko's ellipse provides a visual aid to determining the relationship between variances. If $\text{var}(a)$ is greater than $\text{var}(d)$, the orientation of the ellipse is horizontal. Conversely if $\text{var}(a)$ is less than $\text{var}(d)$, the orientation of the ellipse is vertical.

The Bland-Altman plot for the Grubbs data, complemented by Bartko's ellipse, is depicted in Figure 1.7. The fourth observation is shown to be outside the bounds of the ellipse, indicating that it is a potential outlier.

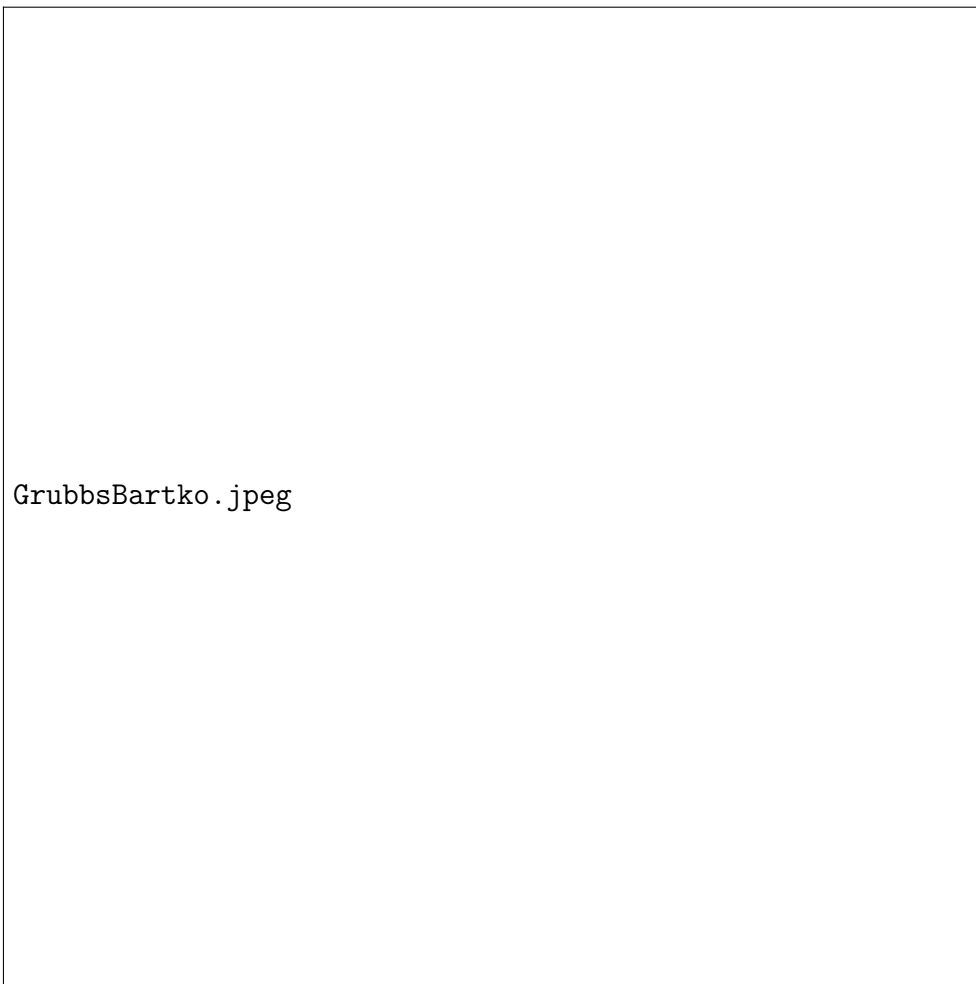



Figure 1.7: Bartko's Ellipse For Grubbs' Data.

The limitations of using bivariate approaches to outlier detection in the Bland-

Altman plot can demonstrated using Bartko's ellipse. A covariate is added to the 'F vs C' comparison that has a difference value equal to the inter-method bias, and an average value that markedly deviates from the rest of the average values in the comparison, i.e. 786. Table 1.8 depicts a 95% confidence ellipse for this manipulated data set. By inspection of the confidence interval, a conclusion would be reached that this extra covariate is an outlier, in spite of the fact that this observation is wholly consistent with the conclusion of the Bland-Altman plot.



GrubbsBartko2.jpeg

Figure 1.8: Bartko's Ellipse For Grubbs' Data, with an extra covariate.

Importantly, outlier classification must be informed by the logic of the data's formulation. In the Bland-Altman plot, the horizontal displacement of any observation

is supported by two independent measurements. Any observation should not be considered an outlier on the basis of a noticeable horizontal displacement from the main cluster, as in the case with the extra covariate. Conversely, the fourth observation, from the original data set, should be considered an outlier, as it has a noticeable vertical displacement from the rest of the observations.

In classifying whether a observation from a univariate data set is an outlier, many formal tests are available, such as the Grubbs test for outliers. In assessing whether a covariate in a Bland-Altman plot is an outlier, this test is useful when applied to the case-wise difference values treated as a univariate data set. The null hypothesis of the Grubbs test procedure is the absence of any outliers in the data set. Conversely, the alternative hypotheses is that there is at least one outlier present.

The test statistic for the Grubbs test (G) is the largest absolute deviation from the sample mean divided by the standard deviation of the differences,

$$G = \max_{i=1,\dots,n} \frac{|d_i - \bar{d}|}{S_d}.$$

For the ‘F vs C’ comparison it is the fourth observation gives rise to the test statistic, $G = 3.64$. The critical value is calculated using Student’s t distribution and the sample size,

$$U = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t_{\alpha/(2n), n-2}^2}{n-2 + t_{\alpha/(2n), n-2}^2}}.$$

For this test $U = 0.75$. The conclusion of this test is that the fourth observation in the ‘F vs C’ comparison is an outlier, with p -value = 0.003, according with the previous result using Bartko’s ellipse.

1.2.5 Inferences on Bland-Altman estimates

? advises on how to calculate confidence intervals for the inter-method bias and limits of agreement. For the inter-method bias, the confidence interval is a simply that of a mean: $\bar{d} \pm t_{(0.5\alpha, n-1)} S_d / \sqrt{n}$. The confidence intervals and standard error for the limits of agreement follow from the variance of the limits of agreement, which is shown to be

$$\text{Var}(LoA) = \left(\frac{1}{n} + \frac{1.96^2}{2(n-1)} \right) s_d^2.$$

If n is sufficiently large this can be following approximation can be used

$$\text{Var}(LoA) \approx 1.71^2 \frac{s_d^2}{n}.$$

Consequently the standard errors of both limits can be approximated as 1.71 times the standard error of the differences.

A 95% confidence interval can be determined, by means of the t distribution with $n-1$ degrees of freedom. However ? comment that such calculations may be ‘somewhat optimistic’ on account of the associated assumptions not being realized.

1.2.6 Formal definition of limits of agreement

? note the similarity of limits of agreement to confidence intervals, but are clear that they are not the same thing. Interestingly, they describe the limits as ‘being like a reference interval’.

Limits of agreement have very similar construction to Shewhart control limits. The Shewhart chart is a well known graphical methodology used in statistical process control. Consequently there is potential for misinterpreting the limits of agreement as if equivalent to Shewhart control limits. Importantly the parameters used to determine the Shewhart limits are not based on any sample used for an analysis, but on the process’s historical values, a key difference with Bland-Altman limits of agreement.

? regards the limits of agreement as a prediction interval for the difference between future measurements with the two methods on a new individual, but states that it does not fit the formal definition of a prediction interval, since the definition does not consider the errors in estimation of the parameters. Prediction intervals, which are often used in regression analysis, are estimates of an interval in which future observations will fall, with a certain probability, given what has already been observed. ? offers an alternative formulation, a 95% prediction interval for the difference

$$\bar{d} \pm t_{(0.975, n-1)} s_d \sqrt{1 + \frac{1}{n}}$$

where n is the number of subjects. Carstensen is careful to consider the effect of the sample size on the interval width, adding that only for 61 or more subjects is there a quantile less than 2.

? offers an alternative description of limits of agreement, this time as tolerance limits. A tolerance interval for a measured quantity is the interval in which a specified fraction of the population's values lie, with a specified level of confidence. ? describes them as a probability interval, and offers a clear description of how they should be used; 'if the absolute limit is less than an acceptable difference d_0 , then the agreement between the two methods is deemed satisfactory'.

The prevalence of contradictory definitions of what limits of agreement strictly are will inevitably attenuate the poor standard of reporting using limits of agreement, as mentioned by ?.

1.2.7 Replicate Measurements

Thus far, the formulation for comparison of two measurement methods is one where one measurement by each method is taken on each subject. Should there be two or more measurements by each methods, these measurement are known as 'replicate

measurements'. ? recommends the use of replicate measurements, but acknowledges the additional computational complexity.

? address this problem by offering two different approaches. The premise of the first approach is that replicate measurements can be treated as independent measurements. The second approach is based upon using the mean of the each group of replicates as a representative value of that group. Using either of these approaches will allow an analyst to estimate the inter method bias.

However, because of the removal of the effects of the replicate measurements error, this would cause the estimation of the standard deviation of the differences to be unduly small. ? propose a correction for this.

? takes issue with the limits of agreement based on mean values of replicate measurements, in that they can only be interpreted as prediction limits for difference between means of repeated measurements by both methods, as opposed to the difference of all measurements. Incorrect conclusions would be caused by such a misinterpretation. ? demonstrates how the limits of agreement calculated using the mean of replicates are 'much too narrow as prediction limits for differences between future single measurements'. This paper also comments that, while treating the replicate measurements as independent will cause a downward bias on the limits of agreement calculation, this method is preferable to the 'mean of replicates' approach.

1.2.8 Bland-Altman correlation test

The approach proposed by ? is a formal test on the Pearson correlation coefficient of case-wise differences and means (ρ_{AD}). According to the authors, this test is equivalent to the 'Pitman Morgan Test'. For the Grubbs data, the correlation coefficient estimate (r_{AD}) is 0.2625, with a 95% confidence interval of (-0.366, 0.726) estimated by Fishers 'r to z' transformation (?). The null hypothesis ($\rho_{AD} = 0$) fail to be rejected. Con-

sequently the null hypothesis of equal variances of each method would also fail to be rejected. There has no been no further mention of this particular test in [1], although [1] refers to Spearman's rank correlation coefficient. [1] comments 'we do not see a place for methods of analysis based on hypothesis testing'. [1] also states that consider structural equation models to be inappropriate.

1.2.9 Identifiability

[1] highlights an important issue regarding using models such as these, the identifiability problem. This comes as a result of there being too many parameters to be estimated. Therefore assumptions about some parameters, or estimators used, must be made so that others can be estimated. For example in literature the variance ratio $\lambda = \frac{\sigma_1^2}{\sigma_2^2}$ must often be assumed to be equal to 1 ([1]). [1] considers methodologies based on two methods with single measurements on each subject as inadequate for a serious study on the measurement characteristics of the methods. This is because there would not be enough data to allow for a meaningful analysis. There is, however, a contrary argument that in many practical settings it is very difficult to get replicate observations when the measurement method requires invasive medical procedure.

[1] offers a formal simultaneous hypothesis test for the mean and variance of two paired data sets. Using simple linear regression of the differences of each pair against the sums, a line is fitted to the model, with estimates for intercept and slope ($\hat{\beta}_0$ and $\hat{\beta}_1$). The null hypothesis of this test is that the mean (μ) and variance (σ^2) of both data sets are equal if the slope and intercept estimates are equal to zero (i.e $\sigma_1^2 = \sigma_2^2$ and $\mu_1 = \mu_2$ if and only if $\beta_0 = \beta_1 = 0$)

A test statistic is then calculated from the regression analysis of variance values ([1]) and is distributed as ' F ' random variable. The degrees of freedom are $\nu_1 = 2$ and $\nu_2 = n - 2$ (where n is the number of pairs). The critical value is chosen for $\alpha\%$

significance with those same degrees of freedom. ? amends this methodology for use in method comparison studies, using the averages of the pairs, as opposed to the sums, and their differences. This approach can facilitate simultaneous usage of test with the Bland-Altman methodology. Bartko’s test statistic take the form:

$$F.test = \frac{(\Sigma d^2) - SSReg}{2MSReg}$$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Averages	1	0.04	0.04	0.74	0.4097
Residuals	10	0.60	0.06		

Table 1.5: Regression ANOVA of case-wise differences and averages for Grubbs Data

For the Grubbs data, $\Sigma d^2 = 5.09$, $SSReg = 0.60$ and $MSreg = 0.06$ Therefore the test statistic is 37.42, with a critical value of 4.10. Hence the means and variance of the Fotobalk and Counter chronometers are assumed to be simultaneously equal.

Importantly, this methodology determines whether there is both inter-method bias and precision present, or alternatively if there is neither present. It has previously been demonstrated that there is a inter-method bias present, but as this procedure does not allow for separate testing, no conclusion can be drawn on the comparative precision of both methods.

1.3 Regression Methods

Conventional regression models are estimated using the ordinary least squares (OLS) technique, and are referred to as ‘Model I regression’ (??). A key feature of Model I models is that the independent variable is assumed to be measured without error.

As often pointed out in several papers (??), this assumption invalidates simple linear regression for use in method comparison studies, as both methods must be assumed to be measured with error.

The use of regression models that assumes the presence of error in both variables X and Y have been proposed for use instead (??). These methodologies are collectively known as ‘Model II regression’. They differ in the method used to estimate the parameters of the regression.

Regression estimates depend on formulation of the model. A formulation with one method considered as the X variable will yield different estimates for a formulation where it is the Y variable. With Model I regression, the models fitted in both cases will entirely different and inconsistent. However with Model II regression, they will be consistent and complementary.

Regression approaches are useful for a making a detailed examination of the biases across the range of measurements, allowing bias to be decomposed into fixed bias and proportional bias. Fixed bias describes the case where one method gives values that are consistently different to the other across the whole range. Proportional bias describes the difference in measurements getting progressively greater, or smaller, across the range of measurements. A measurement method may have either an attendant fixed bias or proportional bias, or both. (?). Determination of these biases shall be discussed in due course.

1.4 Outline of Thesis

Thus the study of method comparison is introduced. The intention of this thesis is to progress the study of method comparison studies, using a statistical method known as Linear mixed effects models. Chapter two shall describe linear mixed effects models,

and how the use of the linear mixed effects models have so far extended to method comparison studies. Implementations of important existing work shall be presented, using the R programming language.

Model diagnostics are an integral component of a complete statistical analysis. In chapter three model diagnostics shall be described in depth, with particular emphasis on linear mixed effects models, further to chapter two.

For the fourth chapter, important linear mixed effects model diagnostic methods shall be extended to method comparison studies, and proposed methods shall be demonstrated on data sets that have become well known in literature on method comparison. The purpose is to both calibrate these methods and to demonstrate applications for them. The last chapter shall focus on robust measures of important parameters such as agreement.

Chapter 2

Linear Mixed effects Models

2.1 Linear Mixed effects Models

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

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

The 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,

$$y = X\beta + Zb + \epsilon \tag{2.1}$$

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 $E(b) = 0$, $E(\epsilon) = 0$ and

$$\text{var}(b\epsilon) = (D00\Sigma$$

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.

2.1.1 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 (??), 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

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

Estimation of the fixed parameters

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

$$\ell(\beta, \theta|y) = -\frac{1}{2} \log |V| - \frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta), \quad (2.2)$$

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

$$(X'V^{-1}X)\beta = X'V^{-1}y. \quad (2.3)$$

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

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

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

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

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

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 β . 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 (??) 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 (??).

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 a 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

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

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

The LME model as a general linear model

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

$$\delta = (\beta b, \quad y_a = (y\psi, \quad T = (XZ0I, \quad \text{and } W = (\Sigma 00D,$$

where ? describe $\psi = 0$ as quasi-data with mean $E(\psi) = b$. 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 $E(\varepsilon) = 0$ and $\text{var}(\varepsilon) = W$, with *both* β and b appearing as fixed parameters. The usefulness of this reformulation of an LME as a general linear model will be revisited.

2.2 Repeated measurements in LME models

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

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

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

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

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

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

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.

2.2.1 Formulation of the response vector

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

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

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

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

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

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

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

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

2.2.2 Decomposition of the response covariance matrix

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

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

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

for repeated measure correlations, such as the compound symmetry and autoregressive structures, as alternative to the identity matrix.

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

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

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

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

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

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

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

2.2.3 Correlation terms

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

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

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

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.

2.3 Using LME for method comparison

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

2.3.1 Correlation

In addition to the variability tests, Roy advises that it is preferable that a correlation of greater than 0.82 exist for two methods to be considered interchangeable. However if two methods fulfil all the other conditions for agreement, failure to comply with this one can be overlooked. Indeed Roy demonstrates that placing undue importance to it can lead to incorrect conclusions. ? remarks that current computer implementations only gives overall correlation coefficients, but not their variances. Consequently it is not possible to carry out inferences based on all overall correlation coefficients.

2.3.2 Variability test 1

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

$$H_0 : d_A = d_B$$

$$H_A : d_A \neq d_B$$

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

2.3.3 Variability test 2

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

$$H_0 : \lambda_A = \lambda_B$$

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

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

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

2.3.4 Variability test 3

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

$$H_0 : \sigma_A = \sigma_B$$

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

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

2.3.5 Demonstration of Roy's testing

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

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

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

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

A likelihood ratio test is perform to compare both candidate models. The log-likelihood of the null model is -2030.7 , and for the alternative model -2030.8 . The test statistic, presented with greater precision than the log-likelihoods, is 0.1592. The p -value is 0.6958. Consequently we fail to reject the null model, and by extension,

conclude that the hypothesis that methods J and S have the same between-subject variability. Thus the second of the criteria is fulfilled.

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

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

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

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

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

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

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

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

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

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

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

2.4 Limits of agreement in LME models

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

? uses LME models to determine the limits of agreement. Between-subject variation for method m is given by d_m^2 and within-subject variation is given by 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 \mathbf{D} and within-subject variability $\mathbf{\Lambda}$ can be presented in matrix form,

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

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

$$\text{var}(y_A - y_B) = 2d^2 + \lambda_A^2 + \lambda_B^2. \quad (2.4)$$

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

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

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

computer output. The limits of agreement are therefore,

$$0.0449 \pm 1.96 \times \sqrt{2 \times 0.0596^2 + 0.0772^2 + 0.0724^2} = (-0.220, 0.309). \quad (2.5)$$

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

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_B^2 + \lambda_A^2 - 2(d_{AB} + \lambda_{AB}) \quad (2.6)$$

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

$$\bar{y}_A - \bar{y}_B \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \quad (2.7)$$

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

2.4.1 Linked replicates

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

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

In most cases, measurements were taken by both method at three different times. In some cases there are either one or two pairs of measurements, hence the data is

unbalanced. ? describes many of the children as being very sick, and with very low oxygen saturations levels. Therefore it must be assumed that a biological change can occur in interim periods, and measurements are not true replicates.

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

Limits of agreement are determined using Roy’s methodology, without adding any additional terms, are found to be consistent with the ‘interaction’ model; (-9.562, 14.504). Roy’s methodology assumes that replicates are linked. However, following Carstensen’s example, an addition interaction term is added to the implementation of Roy’s model to assess the effect, the limits of agreement estimates do not change. However there is a conspicuous difference in within-subject matrices of Roy’s model and the modified model (denoted 1 and 2 respectively);

$$\hat{\Lambda}_1 = \begin{pmatrix} 16.6111 & 16.7111 & 16.727.65 \\ \end{pmatrix} \quad \hat{\Lambda}_2 = \begin{pmatrix} 7.552 & 6.02 & 6.018 & 5.59 \end{pmatrix} (2.8)$$

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

? introduces the Akaike information criterion (*AIC*), a model selection tool based on the likelihood function. Given a data set, candidate models are ranked according to their *AIC* values, with the model having the lowest *AIC* being considered the best fit. Two candidate models can said to be equally good if there is a difference of less than 2 in their *AIC* values.

The Akaike information criterion (*AIC*) for both models are $AIC_1 = 2304.226$ and $AIC_2 = 2306.226$, indicating little difference in models. The *AIC* values for

the Carstensen ‘unlinked’ and ‘linked’ models are 1994.66 and 1955.48 respectively, indicating an improvement by adding the interaction term.

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

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

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

2.5 Implementation in R

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

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

Similarly a variety of structures for the 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 ?, 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.

2.6 Extension of Roy's methodology

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

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

Should an implementation be feasible, further difficulty arises when interpreting the results. The fundamental question is whether two methods have close agreement so as to be interchangeable. When three methods are present in the model, the null hypothesis is that all three methods have the same variability relevant to the respective tests. The outcome of the analysis will either be that all three are interchangeable or that all three are not interchangeable.

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

Roy's methodology is not suitable for the case of single measurements because it

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

2.7 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.