Contents

Chapter 1

Established Techniques for Method Comparison Studies

1.1 Techniques for Simple Design Studies

The issue of whether two methods of measurements are comparable to the extent that they can be used interchangeably with sufficient accuracy and measurement precision is encountered frequency in scientfic research. Published examples of method comparison studies can be found in disciplines as diverse as pharmacology (Ludbrook, 1997), anaesthesia (Myles, 2007), and cardiac imaging methods (Krummenauer et al., 2000)(references).

In the most basic design, items (i.e. people in medical studies) are measures once only by each of two measurement methods. If the recorded measurements by the two instruments differ systematically, a problem of inter-method bias exists. Oftentimes this bias can be mitigated by some technical adjustment or recalibration of the readings. However, if the method variances differ, no comparable adjustment is possible, and a more serious problem exists.

This problem has received significant attention in statistical literature over many decades. Statistical tests for equality of measurements precisions were devised by

Pitman (1939) and Morgan (1939). Grubbs (1948, 1973) formulate a model testing framework for comparing multiple devices.

A graphical tool advocated by Altman and Bland (1983); Bland and Altman (1986) shifted the analysis from concerns over statistical hypothesis testing to concerns of statistical equivalence. Known as the Bland-Altman plot, this technique has become the most popular (and in some cases, obligatory) method of presenting method comparison studies in journals. Dunn (2002) prefers an approach based in measurement error models. Carstensen et al. (2008) extend the technique to replicated deisign using LME frameworks to replicated designs using an LME framework, and supports this work with an R package. Broemeling (2009) lays out a Bayesian strategy.

With some few exceptions, e.g. Hawkins (1978) and Bartko (1994), the issue of outliers and anomalous values have not featured prominently in method comparison literature, a topic that we will particular attention to in a later chapter.

This chapter is organized as follows; firstly a review of the tools used in the analysis of unreplicated deisgns. We then consder ther extension to replicated designs. The LME framework advanced by Carstensen et al. (2008) and Roy (2009) are given special attention.

1.2 Conventional Approaches to Unreplicated Design

Let the random variables Y_1 and Y_2 be distributed bivariate normal with $\mathrm{E}(Y_1) = \mu_1$, $\mathrm{E}(Y_2) = \mu_2$, $\mathrm{var}(Y_1) = \sigma_1^2$, $\mathrm{var}(Y_2) = \sigma_2^2$, and correlation coefficient $-1 < \rho < 1$. Of particular interest are tests of the unconditional marginal hypotheses H': $\mu_1 = \mu_2$ and H'': $\sigma_1^2 = \sigma_2^2$, and tests of the joint hypothesis H^{J}: $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$. The random variables $D = Y_1 - Y_2$ and $S = Y_1 + Y_2$ are bivariate normal with expectations $\mathrm{E}(D) = \mu_D = \mu_1 - \mu_2$ and $\mathrm{E}(S) = \mu_S = \mu_1 + \mu_2$, variances $\mathrm{var}(D) = \sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$ and $\mathrm{var}(S) = \sigma_S^2 = \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2$, and covariance $\mathrm{cov}(D, S) = \sigma_1^2 - \sigma_2^2$. The conditional

distribution of D given S is normal with expectation $\mu_{D|S=s} = \mu_D + [(\sigma_1^2 - \sigma_2^2)/\sigma_S^2](s - \mu_S)$ and variance $\sigma_{D|S}^2 = \sigma_D^2 - (\sigma_1^2 - \sigma_2^2)^2/\sigma_S^2$. These differences and sums are the building blocks of the test procedures: of H$'$, due to Gossett (1908); of H$''$, devised concurrently by Pitman (1939) and Morgan (1939); and of H^J, proposed by Bradley and Blackwood (1989). Notably, the classic test procedure of H$'$ due to Gossett (1908) makes no assumptions about the equality, or otherwise, of the variance parameters σ_1^2 and σ_2^2.

The test procedure for H^J advanced by Bradley and Blackwood (1989) additively decomposes into independent tests of H$''$ and the conditional marginal hypothesis H†: $\mu_1 = \mu_2$, assuming the additional restriction $\sigma_1^2 = \sigma_2^2$. The former test in this decomposition is the Pitman-Morgan procedure referred to above. The latter test in the decomposition is based on the F-ratio with $(1, n - 2)$ degrees-of-freedom, denoted below by F_0^*. Conveniently, all three test procedures can be calculated from the fitted simple linear regression of observed differences on observed sums.

1.2.1 The Pitman-Morgan test

The test of the hypothesis that the variances σ_1^2 and σ_2^2 are equal, which was devised concurrently by Pitman (1939) and Morgan (1939), is based on the correlation of D with S, the coefficient being $\rho_{DS} = (\sigma_1^2 - \sigma_2^2)/(\sigma_D \sigma_S)$, which is zero if, and only if, $\sigma_1^2 = \sigma_2^2$. Consequently a test of H$''$: $\sigma_1^2 = \sigma_2^2$ is equivalent to a test of H: $\rho_{DS} = 0$ and the test statistic is the familiar t-test for a correlation coefficient with $(n - 2)$ degrees-of-freedom:

$$T_{\mathrm{PM}}^* = R\sqrt{\frac{n-2}{1-R^2}},$$

where $R = \sum(D_i - \bar{D})(S_i - \bar{S})/[\sum(D_i - \bar{D})^2 \sum(S_i - \bar{S})^2]^{\frac{1}{2}}$ is the sample correlation coefficient of the n case-wise differences $D_i = Y_{i1} - Y_{i2}$ and sums $S_i = Y_{i1} + Y_{i2}$. Throughout this paper the summation $\sum$ is taken to imply $\sum_{i=1}^n$. The procedure is to reject the hypothesis H$''$ in favour of $\sigma_1^2 \neq \sigma_2^2$ if $|T_{\mathrm{PM}}^*| > t_{\alpha/2,(n-2)\mathrm{df}}$.

1.2.2 The Bradley-Blackwood test

Bradley and Blackwood (1989) write $\mu_{D|S=s} = \mu_D + [(\sigma_1^2 - \sigma_2^2)/\sigma_S^2](s - \mu_S) = \beta_0 + \beta_1 s$ where $\beta_0 = \mu_D - [(\sigma_1^2 - \sigma_2^2)/\sigma_S^2]\mu_S$ and $\beta_1 = (\sigma_1^2 - \sigma_2^2)/\sigma_S^2$. They use this result to propose a test of the joint hypothesis H^J, which is true if, and only if, $\beta_0 = \beta_1 = 0$. Their test procedure follows directly from the theory of linear models (Hogg and Craig, 1995, for example) and is based on the F-ratio

$$F^* = (\frac{n-2}{2})(\frac{\sum D_i^2 - \mathrm{SSE}}{\mathrm{SSE}}) \sim F_{(2, n-2)\mathrm{df}}, \tag{1.1}$$

where SSE is the residual error sum-of-squares from the fitted regression $\hat{D}_i = \hat{\beta}_0 + \hat{\beta}_1 s_i$ of the case-wise differences on the case-wise sums. The procedure is to reject the hypothesis H^J in favour of $\mu_1 \neq \mu_2$ and (or) $\sigma_1^2 \neq \sigma_2^2$ if $F^* > F_{\alpha,(2,n-2)\mathrm{df}}$. The F distribution in (1.1) is valid conditional on S, and since the distribution does not depend on S it is also the unconditional distribution of the test statistic F^*.

Consequently there is no need to make special allowance for the fact that the case-wise sums encountered here are random sums, and not fixed, error-free explanatory variables as regression theory demands. This is the same argument that is generally used to show that t-test for a correlation coefficient is valid, e.g., T^*_{PM} above (Hogg and Craig, 1995, page 499).

Bland-Altman Plots

Altman and Bland (1983) correctly criticised the use the paired differnce reegresion and correlation anlayss for use in method comparison. Their graphical procedure is based on pair-wise differences versus pair-wise averages. This plot is essentially a visual analogue of the quantities underpinning the tests presented previously.

The plot of differences versus average can be obtained by rotating the points in the original scatterplot of X and Y by 45 degrees and rescaling accordingly. This plot, in essence, serves the purpose of a diagnostic plot. From a historical perspective, a similar graphical tool was devised by Tukey several decades earlier Kozak and Wnuk (2014).

We will illustrate the workings of a Bland-Altman plot through a simple example. The data contain in the table below was presented in (Grubbs, 1948). In each of twelve experimental trials, a single round of ammunition was fired from a 155mm artillery piece and its velocity was measured simultaneously (and independently) by three chronographs devices, identified here by the labels 'Fotobalk', 'Counter' and 'Terma'. The Bland-Altman plot for comparing the 'Fotobalk' and 'Counter' methods,

Round	Fotobalk [F]	Counter [C]	Terma [T]	[F-C]	[(F+C)/2]	[F-T]	[(F+T)/2]
1	793.8	794.6	793.2	-0.8	794.2	0.6	793.5
2	793.1	793.9	793.3	-0.8	793.5	-0.2	793.2
3	792.4	793.2	792.6	-0.8	792.8	-0.2	792.5
4	794.0	794.0	793.8	0.0	794.0	0.2	793.9
5	791.4	792.2	791.6	-0.8	791.8	-0.2	791.5
6	792.4	793.1	791.6	-0.7	792.8	0.8	792.0
7	791.7	792.4	791.6	-0.7	792.0	0.1	791.6
8	792.3	792.8	792.4	-0.5	792.5	-0.1	792.3
9	789.6	790.2	788.5	-0.6	789.9	1.1	789.0
10	794.4	795.0	794.7	-0.6	794.7	-0.3	794.5
11	790.9	791.6	791.3	-0.7	791.2	-0.4	791.1
12	793.5	793.8	793.5	-0.3	793.6	0.0	793.5

Table 1.2.1: Fotobalk : Differences and Averages with counter and terma.

which shall henceforth be referred to as the 'F vs C' comparison, is depicted on the right in Figure 1.2.1. The dashed line in the Bland-Altman plot alludes to the inter-method bias between the two methods, estimated by calculating the average of the differences. In the case of Grubbs data the inter-method bias is -0.61 metres per second. By inspection of the plot, one would notice that the differences tend to increase as the averages increase.

The values in the final two columns contain the pairwise differences $d_i = x_i - y_i$ and $a_i = \frac{x_i + y_i}{2}$.

A plot of this quantities is show in figure 1. Also included is a horizontal grey line repesenting th mean of the differences $\bar{d}$. The horizontal dotted lines refer to the
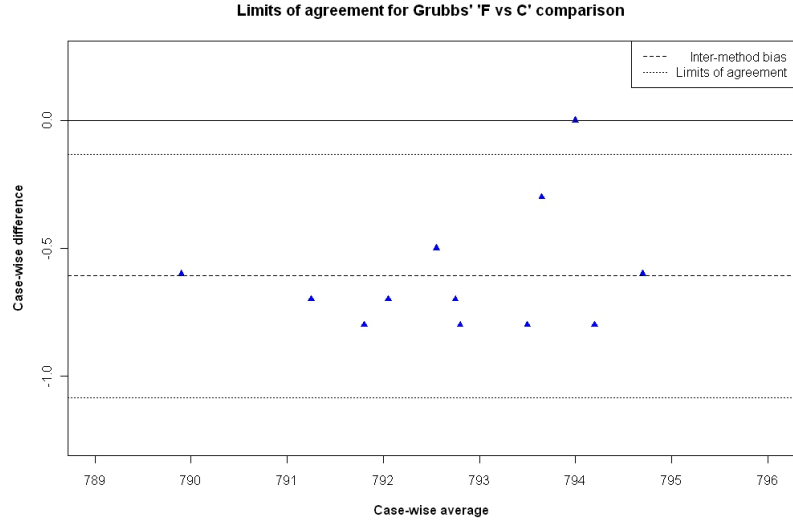
Figure 1.2.1: Bland-Altman plot for Artillery Round Data **?**

limits of agreement and are placed two standard deviations above and below *bard*. The rationale for this plot is that methods showing good agreement would be expected to have values falling predominantly between the limits of agreement.

Bland and Altman (1986) suggested that exact LOAs can be obtained by placing 1.96 in place of the 2 as a multiplier. Bland and Altman (1999) revised this multiplier to be a t value with $n-1$ degrees of freedom for the appropriate coverage. Carstensen et al. (2008) argued that prediction intervals are the appropriate tool for deciding the placment of limits of agreement, and these can be calculated as

$$\bar{d} \pm t_{n-1}. \left[\sqrt{S.E(\bar{d})} \times \frac{n-1}{\sqrt{1+(1/n)}} \right].$$

Altman and Bland (1983) supplement their graphical tool wth a test of the equality of variances, based on the Pitman-Morgan procedure. This test was omitted from their Lancet paper (Bland and Altman, 1986), but was mentioned again in Bland and Altman (1999). In Bland and Altman (1999), they argue that they don't see a role in hypothesis testing in establishing equivalence of measurement methods.

Enhancements proposed by Bland and Altman (1999), such as the use of confidence

7

interval estimates for limits of agreement, have been seldom used in practice.

1.3   Generalized Model Designs

Thus far, the approaches discussed have been based on a simple design, whereby an item is measured by two measurement methods. For method comparison problems, two levels of sophistication exist beyond the simple design.

The first generalization of the design accounts for comparing two instruments when replicate measurements are present. Roy (2009) considers two instruments with replicate measurements a developing a testing framework based on linear mixed effects models.

The second generalization accounts of multiple methods of measurement, e.g. the framework proposed by Grubbs (1973). Carstensen et al. (2008) extends the Bland-Altman framework, also using LME models, to allow for pair-wise comparison of instruments between multiple methods of measurement. Christensen and Blackwood (1993) present a multivariate linear model for method comparison based on Grubbs's model. Their approach generates the multiple correlation test for equality of device variances, and yields a new simultaneous test for equality of variances and biases.

The original Bland-Altman method was developed for two sets of measurements done on one occasion, and so this approach is not suitable for replicate measures data, other than a preliminary exploration of the data. Bland and Altman (1999) addresses the issue of computing LOAs in the presence of replicate measurements, suggesting several computationally simple approaches. When repeated measures data are available, it is desirable to use all the data to compare the two methods.

1.4   LME Models in Method Comparison Studies

When repeated measures data are available for the the computation of the limits of agreement, it is desirable to use all the data to compare the two methods. The clas-

sical Bland-Altman method was developed for two sets of measurements done on one occasion, but is inadequate for replicate measurement data.

Bland and Altman (1999) addressed this issue by suggesting several computationally simple approaches. One approach suggested by Bland and Altman (1999) is to calculate the mean for each method on each subject and use these pairs of means to compare the two methods. Their second approach is to treat each measurement separately.

The estimate of bias will be unaffected using this approach, but the estimate of the standard deviation of the differences will be too small, because of the reduction of the effect of repeated measurement error (Carstensen, 2004; Carstensen et al., 2008). These authors recommends that replicate measurements be used for each method, but recognizes that resulting data are more difficult to analyze. To this end, they recommend the use of LME models as a suitable framework.

Carstensen et al. (2008) extends the well established Bland-Altman methodology for the case of replicate measurements on each item by using LME models, to allow for a more statistically rigourous approach to computing appropriate estimates for the variance of the inter-method bias. As their interest mainly lies in extending the Bland-Altman approach, other formal tests are not considered.

1.4.1 Test For Inter-Method Bias in the LME Frameworks

The presence of an inter-method bias is the source of disagreement between two methods of measurement that is most easily identified. A formal test for this can be implemented by examining the fixed effects of the LME model, which model the inter-method bias. This is common to classical linear model, and interpretation of the results should pose not difficulty to a trained practitioner.

The null hypotheses H_1, that both methods have the same mean, which is tested against the alternative hypothesis K_1, that both methods have different means;

$$H_1 : \mu_1 = \mu_2, \tag{1.2}$$

$$K_1 : \mu_1 \neq \mu_2. \tag{1.3}$$

The inter-method bias and necessary test statistic and $p-$value are presented in computer output.

1.4.2 Limits of Agreement in LME models

Carstensen et al. (2008) recommend a fitted LME model to obtain appropriate estimates for the variance of the inter-method bias. Their interest lies in generalizing the limits-of-agreement (LOA) method developed by Bland and Altman (1986) to take proper cognizance of the replicate measurements. Estimation of repeatability is included in this framework, but other formal tests are not considered. The model is constructed to describe the relationship between a value of measurement and its real value. The non-replicate case is considered first, as it is the context of the Bland-Altman plots. This model assumes that inter-method bias is the only difference between the two methods.

The following model (in the authors own notation) is formulated as follows, where y_{mir} is the rth replicate measurement on subject i with method m. The differences are expressed as $d_i = y_{1i} - y_{2i}$.

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \qquad (1.4)$$

The intercept term α and the $\beta_m \mu_i$ term follow from Dunn (2002), expressing constant and proportional bias respectively, in the presence of a real value μ_i. c_{mi} is a interaction term to account for replicate, and e_{mir} is the residual associated with each observation. Since variances are specific to each method, this model can be fitted separately for each method.

This formulation doesn't require the data set to be balanced, but does require a sufficient number of replicates and measurements to overcome the problem of identifiability. Consequently more than two methods of measurement may be required to carry out the analysis. There is also the assumptions that observations of measurements by particular methods are exchangeable within subjects. The quality of exchangeability means that future samples from a population behaves like earlier samples. For the

10

replicate case, an interaction term c is added to the model, with an associated variance component. All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method.

The measurement y_{mi} by method m on individual i the measurement y_{mir} is the rth replicate measurement on the ith item by the mth method, where $m = 1, 2, \ldots, M$ $i = 1, \ldots, N$, and $r = 1, \ldots, n_i$ is formulated as follows.

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

Here the terms α_m and μ_i represent the fixed effect for method m and a true value for item i respectively. The β_m term follow from Dunn (2002), expressing constant and proportional bias respectively, in the presence of a real value μ_i. We will just consider the case where $\beta = 1$ presently.

c_{mi} is a interaction term to account for replicate, and e_{mir} is the residual associated with each observation. Since variances are specific to each method, this model can be fitted separately for each method.

$$y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + \varepsilon_{mir}. \qquad (1.6)$$

The variation between items for method m is captured by σ_m and the within item variation by τ_m.

All random-effect terms are assumed to be independent. For the case when replicate measurements are assumed to be exchangeable for item i, a_{ir} can be removed. This model describes measurements by m methods, where $m = \{1, 2, 3 \ldots\}$. When the design is balanced and there is no ambiguity we can set $n_i = n$.

The random effect terms comprise an interaction term c_{mi} and the residuals ε_{mir}. The c_{mi} term represent random effect parameters corresponding to the two methods, having $\mathrm{E}(c_{mi}) = 0$ with $\mathrm{Var}(c_{mi}) = \tau_m^2$. All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method. Carstensen et al. (2008) specifies the variance of the interaction terms as being univariate normally distributed. As such, $\mathrm{Cov}(c_{mi}, c_{m'i}) = 0$. All the random

11

effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method. The quality of exchangeability means that future samples from a population behaves like earlier samples.

1.4.3 Linked Replicates

Carstensen et al. (2008) proposes the addition of an random effects term to their model when the replicates are linked. This term is used to describe the 'item by replicate' interaction, which is independent of the methods. This interaction is a source of variability independent of the methods. Therefore failure to account for it will result in variability being wrongly attributed to the methods. Carstensen et al. (2008) demonstrates how to compute the limits of agreement for two methods in the case of linked measurements. As a surplus source of variability is excluded from the computation, the limits of agreement are not unduly wide, which would have been the case if the measurements were treated as true replicates. Failure to take the replication structure into account results in over-estimation of the limits of agreement.

1.4.4 Computation of Limits of Agreement in LME models

Carstensen et al. (2008) proposed a technique to calculate prediction intervals in the presence of replicate measurements, overcoming problems associated with Bland-Altman approach in this regard. Between-subject variation for method m is given by g_m^2 (in the author's notation τ_m^2) and within-subject variation is given by σ_m^2.

Carstensen et al. (2008) states a model where the variation between items for method m is captured by τ_m (our notation g_m^2) and the within-item variation by σ_m. When only two methods are compared, Carstensen et al. (2008) notes that separate estimates of τ_m^2 can not be obtained due to the model over-specification. To overcome this, the assumption of equality, i.e. $\tau_1^2 = \tau_2^2$, is required.

When only two methods are to be compared, separate estimates of τ_m^2 can not be obtained. Instead the average value τ^2 is used. The between-subject variability D and

within-subject variability Λ can be presented in matrix form,

$$G = \begin{pmatrix} g_A^2 & 0 \\ 0 & g_B^2 \end{pmatrix} = \begin{pmatrix} g^2 & 0 \\ 0 & g^2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \sigma_A^2 & 0 \\ 0 & \sigma_B^2 \end{pmatrix}.$$

The variance for method m is $g_m^2 + \sigma_m^2$. 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

$$\text{var}(y_A - y_B) = 2g^2 + \sigma_A^2 + \sigma_B^2. \tag{1.7}$$

Importantly the covariance terms in both variability matrices are zero, so no covariance components are present.

Carstensen et al. (2008) proposes a framework to calculate prediction intervals in the presence of replicate measurements, overcoming problems associated with Bland-Altman approach in this regard. Carstensen et al. (2008) notes that, for $m = 2$, separate estimates of τ_m^2 can not be obtained. To overcome this, the assumption of equality, i.e. $\tau_1^2 = \tau_2^2$ is required, with the limits of agreement therfore as

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

1.4.5   Roy's LME Framework for Method Comparison

Barnhart et al. (2007) sets out three criteria for two methods to be considered in agreement: no significant bias, no difference in the between-subject variabilities, and no significant difference in the within-subject variabilities.

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

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

For the purposes of comparing two methods of measurement, Roy (2009) presents a framework utilizing LME models. This approach 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 (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. 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 as null hypothesis cases.

In addition to computing the inter-method bias, three significance tests are carried out on the respective formulations to make a judgement on whether or not two methods are in agreement. The difference in the models are specifically in how the the D and Σ matrices are constructed, using either an unstructured form or a compound symmetry form. The first model is compared against each of three other models successively. These tests are the pairwise comparison of candidate models, one formulated without constraints, the other with a constraint.

Four candidates models are fitted to the data. These models are similar to one another, but for the imposition of equality constraints. The tests are implemented by fitting a four variants of a specific LME model to the data. For the purpose of comparing models, one of the models acts as a reference model while the three other variant are nested models that introduce equality constraints to serves as null hypothesis cases. The framework uses a linear mixed effects regression fit using a combination of sym-

metric and compound symmetry (CS) correlation structure the variance covariance matrices.

The differences in the candidate 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,

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

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

1.4.6   Model Specification for Roy's Hypotheses Tests

The LME model underpinning Roy's approach can be written

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \tag{1.8}$$

Here β_0 and β_m are fixed-effect terms representing, respectively, a model intercept and an overall effect for method m. The model can be reparameterized by gathering the β terms together into (fixed effect) intercept terms $\alpha_m = \beta_0 + \beta_m$. The b_{1i} and b_{2i} terms are correlated random effect parameters having $\mathrm{E}(b_{mi}) = 0$ with $\mathrm{Var}(b_{mi}) = g_m^2$ and $\mathrm{Cov}(b_{1i}, b_{2i}) = g_{12}$. The random error term for each response is denoted ϵ_{mir} having $\mathrm{E}(\epsilon_{mir}) = 0$, $\mathrm{Var}(\epsilon_{mir}) = \sigma_m^2$, $\mathrm{Cov}(\epsilon_{1ir}, \epsilon_{2ir}) = \sigma_{12}$, $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$ and $\mathrm{Cov}(\epsilon_{1ir}, \epsilon_{2ir'}) = 0$. Additionally these parameter are assumed to have Gaussian distribution. Two methods of measurement are in complete agreement if the null hypotheses $\mathrm{H}_1 \colon \alpha_1 = \alpha_2$ and $\mathrm{H}_2 \colon \sigma_1^2 = \sigma_2^2$ and $\mathrm{H}_3 \colon g_1^2 = g_2^2$ hold simultaneously. Roy (2009) uses a Bonferroni correction to control the familywise error rate for tests of $\{\mathrm{H}_1, \mathrm{H}_2, \mathrm{H}_3\}$ and account for difficulties arising due to multiple testing. Additionally, Roy combines H_2 and H_3 into a single testable hypothesis $\mathrm{H}_4 \colon \omega_1^2 = \omega_2^2$, where $\omega_m^2 = \sigma_m^2 + g_m^2$ represent the overall variability of method m.

1.4.7 Roy's Tests of Variances

Roy (2009) proposes a series of three tests on the variance components of an LME model. For these tests, four candidate models are fitted to the data, each differing by various constraints applied to the variance covariance matrices.

Three tests of hypothesis are provided, appropriate for evaluating the agreement between the two methods of measurement under this sampling scheme.

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.

$$H_2 : \ g_1 = g_2$$
$$K_2 : \ g_1 \neq g_2$$

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

The first test allows of the comparison the begin-subject variability of two methods. Similarly, the second test assesses the within-subject variability of two methods. A third test is a test that compares the overall variability of the two methods. Other important aspects of the method comparison study are consequent. The limits of agreement are computed using the results of the first model.

$$H_3 : \sigma_1^2 = \sigma_2^2$$
$$K_3 : \sigma_1^2 \neq \sigma_2^2$$

Variability Test 2

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

$$H_4: \quad \sigma_1 = \sigma_2$$

$$K_4: \quad \sigma_1 \neq \sigma_2$$

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

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

Variability Test 3

Roy also integrates H$_2$ and H$_3$ into a single testable hypothesis H$_4$: $\omega_1^2 = \omega_2^2$, where $\omega_m^2 = \sigma_m^2 + g_m^2$ represent the overall variability of method m. Roy (2009) further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal. An examination of this topic is useful because a method for computing Limits of Agreement follows from here.

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

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

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

Two methods can be considered to be in agreement if criteria based upon these techniques are met. Additionally Roy makes reference to the overall correlation coefficient of the two methods, which is determinable from variance estimates.

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

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

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

$$\text{Block } \Omega_i = \hat{D} + \hat{\Sigma} \tag{1.9}$$

Overall variability between the two methods (Ω) is sum of between-subject (D) and within-subject variability (Σ), Roy (2009) denotes the overall variability as Block - Ω_i. The overall variation for methods 1 and 2 are given by

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

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

$$H_0 : \ \omega_1 = \omega_2$$

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

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

1.4.8   Computing Limits of Agreement Using Roy's Model

Roy (2009) has demonstrated a method whereby g_A^2 and g_B^2 can be estimated separately. Also covariance terms are present in both D and Σ. Using Roy's approach, the variance of case-wise difference in measurements can be determined from Block-Ω_i. Hence limits of agreement can be computed. The computation of the limits of agreement require that the variance of the difference of measurements. This variance is easily computable from the estimate of the Block - Ω_i matrix. The variance of differences is easily computable from the variance estimates in the Block - Ω_i matrix, i.e.

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

Lack of agreement can arise if there is a disagreement in overall variabilities.

The limits of agreement computed by Roy's method are derived from the variance covariance matrix for overall variability. This matrix is the sum of the between subject VC matrix and the within-subject Variance Covariance matrix. For Carstensen's 'fat' data, the limits of agreement computed using Roy's method are consistent with the estimates given by Carstensen et al. (2008); $0.044884 \pm 1.96 \times 0.1373979 = (-0.224, 0.314)$.

1.4.9   Role of Covariance Estimates

In many cases the limits of agreement derived from this method accord with those to Roy's model. However, in other cases dissimilarities emerge. An explanation for this differences can be found by considering how the respective models account for covariance in the observations.

Specifying the relevant terms using a bivariate normal distribution, Roy's model allows for both between-method and within-method covariance. Carstensen et al. (2008) formulate a model whereby random effects have univariate normal distribution, and no allowance is made for correlation between observations.

In contrast to Roy's model, Carstensen's model requires that commonly used assumptions be applied, specifically that the off-diagonal elements of the between-item

and within-item variability matrices are zero. By extension the overall variability off-diagonal elements are also zero. Therefore the variance covariance matrices for between-item and within-item variability are respectively.

$$D = \begin{pmatrix} g_2^1 & 0 \\ 0 & g_2^2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_2^1 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

As a consequence, Carstensen's method does not allow for a formal test of the between-item variability.

In cases where there is negligible covariance between methods, the limits of agreement computed using Roy's model accord with those computed using model described by Carstensen et al. (2008).

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

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

There is a substantial difference in the number of fixed parameters used by the respective models; the model in Roy (2009) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items N, whereas the model using the Carstensen Model requires $N + 2$ fixed effects. Allocating fixed effects to each item i using Carstensen's model accords with earlier work on comparing methods of measurement, such as Grubbs (1948). However allocation of fixed effects in ANOVA models suggests that the group of items is itself of particular interest, rather than as a representative sample used of the overall population. However this approach seems

contrary to the purpose of LoAs as a prediction interval for a population of items. Conversely, Roy (2009) uses a more intuitive approach, treating the observations as a random sample population, and allocating random effects accordingly.

1.4.10   Comparing MCS Approaches

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

Allocating fixed effects to each item i by (1.6) accords with earlier work on comparing methods of measurement, such as Grubbs (1948). However allocation of fixed effects in ANOVA models suggests that the group of items is itself of particular interest, rather than as a representative sample used of the overall population. However this approach seems contrary to the purpose of LOAs as a prediction interval for a population of items. Conversely, Roy (2009) uses a more intuitive approach, treating the observations as a random sample population, and allocating random effects accordingly.

Bibliography

Altman, D. and J. Bland (1983). Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society. Series D (The Statistician) 32*(3), 307–317.

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

Bartko, J. (1994). Measures of agreement: A single procedure. *Statistics in Medicine 13*, 737–745.

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.

Bradley, E. L. and L. G. Blackwood (1989). Comparing paired data: A simultaneous test for means and variances. *The American Statistician 43*(4), 234–235.

Broemeling, L. D. (2009). *Bayesian methods for measures of agreement.* CRC Press.

Carstensen, B. (2004). Comparing and predicting between several methods of measurement. *Biostatistics 5*(3), 399–413.

Carstensen, B. (2010). Comparing methods of measurement: Extending the loa by regression. *Statistics in medicine 29*(3), 401–410.

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).

Christensen, R. and L. G. Blackwood (1993). Tests for precision and accuracy of multiple measuring devices. *Technometrics 35* (4), 411–420.

Dunn, G. (2002). *Statistical Evaluation of Measurement Error* (Second ed.). Stanford: American Mathematical Society and Digital Press.

Gossett, W. S. G. (1908). The probable error of a mean. *Biometrika 6* (1), 1–25.

Grubbs, F. (1948). On estimating precision of measuring instruments and product variability. *Journal of the American Statistical Association 43*, 243–264.

Grubbs, F. (1973). Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. *Technometrics 15* (1), 53–66.

Hawkins, D. M. (1978). Analysis of three tests for one or two outliers. *Statistica Neerlandica 32* (3), 137–148.

Hogg, R. V. and A. T. Craig (1995). *Introduction to mathematical statistics.(5"" edition).* Upper Saddle River, New Jersey: Prentice Hall.

Kozak, M. and A. Wnuk (2014). Including the tukey mean-difference (bland–altman) plot in a statistics course. *Teaching Statistics 36* (3), 83–87.

Krummenauer, F., I. Genevriere, and U. Nixdorff (2000). The biometrical comparison of cardiac imaging methods. *Computer Methods and Programs in Biomedicine 62*, 21–34.

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.

Ludbrook, J. (1997). Comparing methods of measurement. *Clinical and Experimental Pharmacology and Physiology 24*, 193–203.

Morgan, W. A. (1939). A test for the signicance of the difference between two variances in a sample from a normal bivariate population. *Biometrika 31*, 13–19.

Myles, P. (2007). Using the Bland Altman method to measure agreement with repeated measures. *British Journal of Anaesthesia 99*(3), 309–311.

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika 31*, 9–12.

Roy, A. (2009). 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.