

## 0.1 Roy's method

Roy proposes a novel method using the LME model with Kronecker product covariance structure in a doubly multivariate set-up to assess the agreement between a new method and an established method with unbalanced data and with unequal replications for different subjects (?).

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

## 0.2 Hypothesis Testing

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

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

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

# 1 Introduction

? uses an approach based on linear mixed effects (LME) models for the purpose of comparing the agreement between two methods of measurement, where replicate measurements on items (often individuals) by both methods are available. She provides

three tests of hypothesis appropriate for evaluating the agreement between the two methods of measurement under this sampling scheme. These tests consider null hypotheses that assume: absence of inter-method bias; equality of between-subject variabilities of the two methods; equality of within-subject variabilities of the two methods. By inter-method bias we mean that a systematic difference exists between observations recorded by the two methods. Differences in between-subject variabilities of the two methods arise when one method is yielding average response levels for individuals that are more variable than the average response levels for the same sample of individuals taken by the other method. Differences in within-subject variabilities of the two methods arise when one method is yielding responses for an individual that are more variable than the responses for this same individual taken by the other method. The two methods of measurement can be considered to agree, and subsequently can be used interchangeably, if all three null hypotheses are true.

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

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \quad (1)$$

Here  $\beta_0$  and  $\beta_m$  are fixed-effect terms representing, respectively, a model intercept and an overall effect for method  $m$ . The model can be reparameterized by gathering the  $\beta$  terms together into (fixed effect) intercept terms  $\alpha_m = \beta_0 + \beta_m$ . The  $b_{1i}$  and  $b_{2i}$  terms are correlated random effect parameters having  $E(b_{mi}) = 0$  with  $\text{Var}(b_{mi}) = g_m^2$  and  $\text{Cov}(b_{1i}, b_{2i}) = g_{12}$ . The random error term for each response is denoted  $\epsilon_{mir}$  having  $E(\epsilon_{mir}) = 0$ ,  $\text{Var}(\epsilon_{mir}) = \sigma_m^2$ ,  $\text{Cov}(\epsilon_{1ir}, \epsilon_{2ir}) = \sigma_{12}$ ,  $\text{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$  and  $\text{Cov}(\epsilon_{1ir}, \epsilon_{2ir'}) = 0$ . Additionally these parameters are assumed to have Gaussian distribution. Two methods of measurement are in complete agreement if the null hypotheses  $H_1: \alpha_1 = \alpha_2$  and  $H_2: \sigma_1^2 = \sigma_2^2$  and  $H_3: g_1^2 = g_2^2$  hold simultaneously. ? uses a Bonferroni correction to control the familywise error rate for tests of  $\{H_1, H_2, H_3\}$  and account

for difficulties arising due to multiple testing. Additionally, Roy combines  $H_2$  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  $m$ .

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

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

$$y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + \varepsilon_{mir}. \quad (2)$$

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

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

? notes that, for  $m = 2$ , separate estimates of  $\tau_m^2$  can not be obtained. To overcome this, the assumption of equality, i.e.  $\tau_1^2 = \tau_2^2$  is required.

There is a substantial difference in the number of fixed parameters used by the respective models; the model in (1) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items  $N$ , whereas the model in (2) requires  $N + 2$  fixed effects.

Allocating fixed effects to each item  $i$  by (2) accords with earlier work on comparing methods of measurement, such as ?. However allocation of fixed effects in ANOVA models suggests that the group of items is itself of particular interest, rather than as a representative sample used of the overall population. However this approach seems contrary to the purpose of LOAs as a prediction interval for a population of items. Conversely, ? uses a more intuitive approach, treating the observations as a random sample population, and allocating random effects accordingly.

## 2 Roy's Hypotheses Tests

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

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

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

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

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

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

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

where  $\sigma_1^2$  and  $\sigma_2^2$  are the within-subject variances of the respective methods, and  $\sigma_{12}$  is the within-item covariance between the two methods. The within-item variance covariance matrix  $\boldsymbol{\Sigma}$  is assumed to be the same for all replications. Computational analysis of linear mixed effects models allow for the explicit analysis of both  $\mathbf{G}$  and  $\mathbf{R}_i$ .

For expository purposes consider the case where each item provides three replicate measurements by each method. In matrix form the model has the structure

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

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

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

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

$$\begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \quad (3)$$

## 2.1 Roy's hypothesis tests

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

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

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

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

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

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

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

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

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

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

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

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

These tests are complemented by the ability to the overall correlation coefficient of the two methods, which is determinable from variance estimates. Two methods can be considered to be in agreement if criteria based upon these tests are met. Inference

for inter-method bias follows from well-established methods and, as such, will only be noted when describing examples.

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

## 2.2 Variance Covariance Matrices

Under Roy's model, random effects are defined using a bivariate normal distribution. Consequently, the variance-covariance structures can be described using  $2 \times 2$  matrices. A discussion of the various structures a variance-covariance matrix can be specified under is required before progressing. The following structures are relevant: the identity structure, the compound symmetric structure and the symmetric structure.

The identity structure is simply an abstraction of the identity matrix. The compound symmetric structure and symmetric structure can be described with reference to the following matrix (here in the context of the overall covariance Block- $\mathbf{\Omega}_i$ , but equally applicable to the component variabilities  $\mathbf{G}$  and  $\mathbf{\Sigma}$ );

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

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