ROBUST ESTIMATION OF VARIANCE COMPONENTS

by

Werner A. Stahel
Swiss Federal Institute of Technology (ETH), Zurich,
Switzerland
and
Alan Welsh
Australian National University, Canberra, Australia

Research Report No. 69
Sep 24, 1992

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

ROBUST ESTIMATION OF VARIANCE COMPONENTS

Werner A. Stahel
Swiss Federal Institute of Technology (ETH), Zurich, Switzerland
and
Alan Welsh
Australian National University, Canberra, Australia
Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

Sep 24, 1992

Abstract

This paper deals with the problem of estimating the variances in the simple balanced two component model for a one-way design. We describe the classical approaches to estimating variance components and then review the extant proposals for robust approaches, exposing relationships between the various proposals. Some new robust procedures are proposed.

A small simulation study shows that (i) the possible gains of robust procedures over classical ones for contaminated situations are much greater that in location problems; (ii) the proposals of Fellner (1986) and Rocke (1983) work well, but modifications can clearly improve their efficiency for uncontaminated cases.

KEY WORDS: Hierarchical Models, Model II, Random Effects.

1 Introduction

Linear models with multiple stochastic error components are widely used in statistics. Traditionally, their domain of application has included survey sampling (Yates and Zacopancy 1935, Cochran 1939), the analysis of designed experiments (Yates 1940, Rao 1947), genetics (Fairfield, Smith 1936, Henderson 1950) and industrial problems (Brownlee 1953). Our purpose in this paper is to review some recently proposed methods for the robust estimation of the variances of these components (known as variance components) and to make some new proposals.

Some aspects of general variance component models (such as identifiability) turn out to be quite delicate (see for example the review paper of Rao and Kleffe 1980), so for simplicity and clarity we restrict ourselves throughout this paper to the simplest possible variance component model, namely the balanced two-component model for a one-way design. This model enables us to explore many of the issues which arise in the application of general variance component models, but is simple and concrete enough for us to gain insight into the nature of proposed estimation methods. It is of course a basic requirement of any general proposal that it should perform well and exhibit its intended properties in the simplest possible case.

In the balanced two-component model we have n observations y_{ij} on each of g groups which we can write as

$$y_{ij} = \mu + a_i + e_{ij} \ , \quad j = 1, \ldots, m \ , \quad i = 1, \ldots, g \ , \tag{1.1}$$

where the unobservable random variables $\{a_i\}$ and $\{e_{ij}\}$ are independent and have independent $N(0, \sigma_a^2)$ and $N(0, \sigma_e^2)$ distributions, respectively. It follows that the y_{ij} are jointly normally distributed with mean μ and

$$\mathrm{Cov}(y_{ij}, y_{k\ell}) = \begin{cases} \sigma_e^2 + \sigma_a^2 & i = k, \ j = \ell \\ \sigma_a^2 & i = k, \ j \neq \ell \\ 0 & i \neq k \ . \end{cases} \tag{1.2}$$

The unknown parameters are μ, σ_e^2, σ_a^2, the last two of which are the variance components.

An equivalent multivariate formulation of the model is obtained by supposing that we make g m-vector observations y_i which we can write as

$$y_i = \mu 1_m + a_i 1_m + e_i \ , \quad i = 1, \ldots, g \ , \tag{1.3}$$

where 1_m is an m-vector of ones and the unobservable random variables $\{a_i\}$ and $\{e_i\}$ are independent and have independent $N(0, \sigma_a^2)$ and $N_m(0, \sigma_e^2 I_m)$

distributions, respectively. The y_i are then independent $N_m(\mu 1_m, \Sigma)$ random variables with

$$\Sigma = \sigma_e^2 I_m + \sigma_a^2 1_m 1_m^T \ .$$

Finally, we can also write $y = (y_1^T, \ldots, y_g^T)^T$ as

$$y = \mu 1_{mg} + Za + e \ , \quad Z = \begin{pmatrix} 1_m & & 0 \\ & \ddots & \\ 0 & & 1_m \end{pmatrix}_{mg \times g} \ , \qquad (1.5)$$

where a and e are independent $N_g(0, \sigma_a^2 I_g)$ and $N_{mg}(0, \sigma_e^2 I_{mg})$ random variables, respectively. Then, $y \sim N_{mg}(\mu 1_{mg}, V)$, where

$$V = \sigma_e^2 I_{mg} + \sigma_a^2 Z Z^T \ .$$

These two multivariate formulations are convenient for motivating and describing some of the proposed estimation methods and will be used interchangeably with the analysis of variance formulation.

The model is often formulated purely as a model for the mean and covariance structure of y without the imposition of normality. The classical procedures described in the next section then address the distribution–free or "nonparametric" problem of estimating $\mu = E(y_{ij})$ and σ_a^2 and σ_e^2 in (1.2) regardless of the distributions of a_i and e_{ij}. Many properties of these procedures only depend on the existence of fourth moments (Harville 1977, Westfall 1988).

In distributional robustness theory, there are two points of view which usually lead to similar results. The first entertains the idea that there are genuine ("good") observations which follow a "central model". In our case, this will be the model assumed above, with the assumption of normality. These good observations are "contaminated" by a proportion ε of observations which come from a different source. The problem is to estimate the parameters of the central model. Even asymptotically ($g \to \infty$ in our case), there is an "unavoidable bias", since the same contaminated distribution might result from contaminating different members of the central parametric model – the parameters are not identifiable given only the distribution of the observations. Nevertheless, the bias with respect to a fixed central distribution can be studied as a function of the contamination. Its supremum over a set of contaminations (a neighborhood of the central model) can be minimised, leading to minimax–bias estimators (Martin and Zamar, 1989).

Simpler results are obtained if the maximal bias is approximated using the influence function. Then, minimisation leads to the most B-robust estimators. Furthermore, the random variation can be taken into account, either exclusively, as in minimax asymptotic variance estimation (Huber 1964), or in conjuntion with the approximate bias, leading to optimal B-robust estimation. More details can be found in Hampel et al. (1986, chap. 8.2d). Since the supremum of the bias over contamination "neighborhoods" is infinite for classical procedures, they are non-robust. Also, the fourth moment does not exist for all distributions in such neighborhoods.

In the second approach, a weakly continuous functional of the distribution is defined as the quantity to be estimated. Consistent estimators must be asymptotically equivalent to applying this functional to the empirical distribution function. Robustness is then a property of the posed problem rather than the statistical procedure. The functional should be chosen in such a way that it varies little for distributions which are near each other in the sense of a suitable metric. Similar optimality problems can be posed as for the first approach. For details, see Rieder (1991).

The various approaches to robustness are typically simplified by an asymptotic treatment. In the present context, we allow at least $g \to \infty$. Notice that if we regard the group effects as fixed, it is desirable to keep g (and hence the number of parameters) small, but if we want to model the distribution of the group effects, we need a larger number of groups. In particular, for those data sets for which g is small, precise estimation of σ_a^2 is impossible.

In our case, both the random errors and the random effects can be contaminated. In principle, robust procedures should be able to withstand contamination in both components (Youden 1975, Rocke 1983, Fellner 1986). Other important departures from the underlying model, such as non-additivity or lack of independence of errors and/or random effects, will not be considered here.

Variance component models are based on an additive decomposition of variability into several components. Rocke (1983) states that this decomposition is a particular property of the variance which is not shared by other measures of spread. While this gives rise to a slight conceptual complication in the second approach to robustness discussed above, the first one is not affected as long as the central model admits this decomposition.

Note that a decomposition of spread measures is indeed possible for all stable distributions, not only the normal. The suitable, additive measures are σ_e^α and σ_a^α if the e_{ij} and the a_i are (symmetric) stable with the same shape parameter α and the scale parameters σ_e and σ_a, respectively. Such a model may well serve as the central model in the sense just described.

It is worth mentioning that applied practice does recognise the deleterious effect of outliers on variance component estimation. The problem is traditionally addressed by ad hoc deletion methods followed by the application of a classical estimation method to the "clean" data (Youden 1975, Rocke 1983). This approach is difficult to analyse theoretically (so consistency adjustments are impractical) and has been found by Rocke (1983) to produce estimates with large negative biases at the model.

In this paper, we review the known proposals for robust estimates of variance components in the framework of the simplest case (1.1) and add some new ones (Section 3), after describing the classical approaches on which they rely (Section 2).

In general, there may be interest in estimating the random effects (a_i in our case) rather than their spread only. Applications include the determination of breeding values in animal breeding and risks for individual contracts in insurance. Such estimates arise as by-products in most procedures for variance components estimation. We shall not treat this problem in any depth here.

2 Classical estimation procedures

Notation. Throughout the paper, let $\bar{y}_{i\cdot} = m^{-1}\Sigma_{j=1}^{m} y_{ij}$ denote the mean of the ith group and $\bar{y}_{\cdot\cdot} = (mg)^{-1}\Sigma_{j=1}^{m}\Sigma_{i=1}^{g} y_{ij}$ denote the grand mean. We apply the same notation to the errors and the random effects.

Quadratic forms. The oldest class of estimators of the variance components σ_a^2 and σ_e^2 consists of estimators based on location invariant quadratic functions of y. Specifically, the estimators are

$$\widehat{\sigma}_e^2 = y^T Q_e y \ , \quad \widehat{\sigma}_a^2 = y^T Q_a y$$

with suitable Q_e and Q_a. The requirement of translation equivariance,

$$\widehat{\sigma}_e^2(y) = \widehat{\sigma}_e^2(y + \text{const} \cdot 1_{gm}) \ ,$$

leads to

$$Q_e 1_{gm} = 0 \ , \tag{2.1}$$

and similarly for Q_a. Unbiased estimators satisfy

$$E(y^T Q_e y) = \sigma_e^2 \ , \quad E(y^T Q_a y) = \sigma_a^2 \ . \tag{2.2}$$

The problem then is to specify a particular Q_a and Q_e satisfying (2.1) and (2.2).

The simplest estimators of σ_e^2 and σ_a^2 are obtained by equating the mean squares within and between groups from the usual analysis of variance table to their expectations. We obtain

$$\widehat{\sigma}_e^2 = \frac{1}{g(m-1)} \sum_{i=1}^{g} \sum_{j=1}^{m} (y_{ij} - \bar{y}_{i\cdot})^2 \ , \tag{2.3}$$

$$\widehat{\sigma}_a^2 = \frac{1}{g-1} \sum_{i=1}^{g} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 - \frac{1}{m} \widehat{\sigma}_e^2 \ . \tag{2.4}$$

This is equivalent to choosing

$$\begin{aligned} Q_e &= \tfrac{1}{g(m-1)} (I_{mg} - Z(Z^T Z)^{-1} Z^T) \ , \\ Q_a &= \tfrac{1}{m} \{ \tfrac{1}{g-1} (Z(Z^T Z)^{-1} Z^T - \tfrac{1}{mg} 1_{mg} 1_{mg}^T) - Q_e \} \ . \end{aligned}$$

While Q_e is non-negative definite, Q_a is not, so $\widehat{\sigma}_a^2$ can be negative even though $\sigma_a^2 > 0$.

In the first systematic study of variance component estimators Henderson (1953) suggested choices of Q_a and Q_e based on the mean squares associated with alternative analysis of variance tables. In particular, for the balanced two component problem his recommendation leads to this simple estimator.

MINQE. The simplicity of Henderson's (1953) approach is mitigated by its arbitrariness. This is readily apparent in unbalanced problems for which the choice of Q_a and Q_e is more difficult and for which other approaches seem to perform better (Harville, 1977). Probably the most interesting and elegant alternative approach to choosing Q_a and Q_e is minimum norm quadratic estimation (MINQE) approach proposed by Rao (1970) and subsequently developed by him and others (see Rao and Kleffe, 1980, 1988). Consider for simplicity the problem of estimating σ_a^2. The basic idea is that if a and e were known, we would estimate σ_a^2 by $a^T a/g$ or, equivalently, by $(a^T, e^T) M_a (a^T, e^T)^T$, where

$$M_a = \begin{pmatrix} g^{-1} I_g & 0 \\ 0 & 0 \end{pmatrix} .$$

For any Q satisfying $Q 1_{mg} = 0$, (1.5) leads to

$$y^T Q y = (a^T, e^T)(Z, I_{mg})^T Q (Z, I_{mg})(a^T, e^T)^T .$$

Therefore,

$$\begin{aligned} & y^T Q_a y - a^T a/g \\ & \quad = (a^T, e^T)\{(Z, I_{mg})^T Q_a (Z, I_{mg}) - M_a\}(a^T, e^T)^T \end{aligned} .$$

If σ_a^2 and σ_e^2 were known, this could be expressed in terms of standardised random variables as

$$y^T Q y - a^T a/g = (a^T/\sigma_a, e^T/\sigma_e) C (a^T/\sigma_a, e^T/\sigma_e)^T$$

with

$$C = (\sigma_a Z, \sigma_e I_{mg})^T Q_a (\sigma_a Z, \sigma_e I_{mg}) - \begin{pmatrix} \sigma_a^2 I_g/g & 0 \\ 0 & 0 \end{pmatrix} .$$

The MINQE principle consists of minimising a suitable norm of C to obtain Q_a. Since σ_a and σ_e are not known, preliminary estimates must be used to calculate C. If we use the Euclidean norm and impose (2.1) and (2.2), the MINQE estimators are closely related to the locally minimum variance unbiased estimators (MIVQUE) of LaMotte (1973) and Rao (1971 a,b).

Non-negative definiteness constraints can be incorporated into MINQE. However, these constraints complicate the optimisation problem, see Pukelsheim (1981) and Rao and Kleffe (1988). It is therefore common to truncate a negative estimate to zero.

Finally, Zielinski (1986) proposed modifying the MIVQUE estimator to obtain a more stable estimator. Instead of choosing the quadratic form with minimum variance, he suggested using the quadratic form for which the variance changes as little as possible as the kurtosis of the distribution of the random effects varies over pre-specified intervals.

Maximum likelihood. According to formulation (1.3), the log likelihood can be written as $\sum_i \ell(y_i; \sigma_a^2, \sigma_e^2, \mu)$. To make derivations more transparent, we write

$$\tau_a = \sigma_a^2 \ , \quad \tau_e = \sigma_e^2$$

and maximise with respect to μ, τ_e, and

$$\tau_m = \tau_a + \tau_e/m \ .$$

The summand ℓ is

$$\begin{aligned} 2\ell(y_i; \tau_m, \tau_e, \mu) &= -\log|\underset{\sim}{\Sigma}| - (y_i - 1_m\mu)^T \underset{\sim}{\Sigma}^{-1}(y_i - 1_m\mu) + \text{const} \\ &= -(m-1)\log\tau_e - \log(m\tau_m) - (y_i - 1_m\mu)^T(y_i - 1_m\mu)/\tau_e \\ &\quad -(1/\tau_m - m/\tau_e)(\bar{y}_{i\cdot} - \mu)^2 + \text{const} \ . \end{aligned} \tag{2.5}$$

Here we have used the method of diagonal expansion (e.g. Searle 1982, p.106) to evaluate the determinant and the well known inversion formula for $\underset{\sim}{\Sigma}$ (ibid., p.132). It is interesting to note that the log–likelihood depends both on the observations and on the normalised group means. Provided the maximum is not attained on the boundary, the maximum likelihood estimators $\hat{\tau}_m$, $\hat{\tau}_e$, $\hat{\mu}$ are a solution of

$$\sum_{i=1}^{g} s^{(\tau_m)}(y_i; \tau_m, \tau_e, \mu) = 0 \tag{2.6}$$

$$\sum_{i=1}^{g} s^{(\tau_e)}(y_i; \tau_m, \tau_e, \mu) = 0 \tag{2.7}$$

$$\sum_{i=1}^{g} s^{(\mu)}(y_i; \tau_m, \tau_e, \mu) = 0 \ , \tag{2.8}$$

where $s^{(\tau_m)}$, $s^{(\tau_e)}$, and $s^{(\mu)}$ are the Fisher scores, i.e., the derivatives of ℓ

with respect to τ_m, τ_e, and μ, respectively,

$$\begin{aligned} s^{(\tau_m)}(y_i;\tau_m,\tau_e,\mu) &= \frac{1}{2\tau_m}[(\bar{y}_{i\cdot}-\mu)^2/\tau_m - 1] \\ s^{(\tau_e)}(y_i;\tau_m,\tau_e,\mu) &= \frac{1}{2\tau_e}[(y_i - 1_m\bar{y}_{i\cdot})^T(y_i - 1_m\bar{y}_{i\cdot})/\tau_e - (m-1)] \\ s^{(\mu)}(y_i;\tau_m,\tau_e,\mu) &= \frac{1}{\tau_m}(\overline{y}_{i\cdot}-\mu) \ . \end{aligned} \tag{2.9}$$

[For the second line, we have used $(y_i - 1_m\mu)^T(y_i - 1_m\mu) - m(\bar{y}_{i\cdot}-\mu)^2 = (y - 1_m\bar{y}_{i\cdot})^T(y - 1_m\bar{y}_{i\cdot})$.] It follows that

$$\begin{aligned} \widehat{\mu} &= \bar{y}_{\cdot\cdot} \\ \widehat{\tau}_m &= \tfrac{1}{g}\sum_{i=1}^{g}(\bar{y}_{i\cdot}-\widehat{\mu})^2 \\ \widehat{\tau}_e &= \tfrac{1}{g(m-1)}\sum_{i=1}^{g}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})^2 \ , \end{aligned} \tag{2.10}$$

and this coincides with (2.4) up to a factor $\frac{g-1}{g}$ for $\widehat{\tau}_m$.

Let us note in passing that $\widehat{\tau}_m$, $\widehat{\tau}_e$, and $\widehat{\mu}$ are independent, and that the information matrix is

$$J(\widehat{\tau}_m,\widehat{\tau}_e,\widehat{\mu};\tau_m,\tau_e,\mu) = g\cdot\ \mathrm{diag}\,[\,1/(2\tau_m^2),\ (m-1)/(2\tau_e^2),\ \tau_m^{-1}\,]\ . \tag{2.11}$$

If $\widehat{\tau}_m - \widehat{\tau}_e/m < 0$, then the maximum of the likelihood occurs at the boundary $\tau_a = 0$. Maximisation with respect to τ_e and μ can again be done by equating derivatives to zero, observing that $\tau_m = \tau_e/m$. This leads to $\widehat{\mu} = \bar{y}_{\cdot\cdot}$ and

$$\widehat{\tau}_e = \tfrac{1}{gm}\sum_{i=1}^{g}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{\cdot\cdot})^2\ .$$

Reduced Maximum Likelihood. Maximum likelihood estimators of variances are often biased in finite samples but this bias can be reduced by simple adjustment. A heuristic argument is helpful here. Note that the expectations of the left hand sides of (2.6), (2.7), and (2.8) are zero when the true parameters are used, and the maximum likelihood estimate is obtained by equating these zero expectations with their empirical values. Substituting μ by $\bar{y}_{..}$ in (2.6) changes the expectation of the left hand side to τ_m. Equating the empirical value to this expectation instead of zero yields $\widehat{\tau}_m = \frac{1}{g-1}\Sigma_i(\bar{y}_{i.} - \bar{y}_{..})^2$. This idea is one way to interpret the method proposed by Patterson and Thompson (1971) in a more general context under the name of reduced maximum likelihood estimate (REML). It may alternatively be obtained by differentiating the marginal likelihood for τ_m and τ_e (or σ_a^2 and σ_e^2), i.e., the likelihood of any $mg - 1$ components of the mg-vector $y - 1_{mg}\bar{y}_{..}$. When $\widehat{\tau}_m - \widehat{\tau}_e/m < 0$, the REML estimate of τ_e is

$$\widehat{\tau}_e = \frac{1}{g(m-1)} \sum_{i=1}^{g} \sum_{j=1}^{m} (y_{ij} - \bar{y}_{..})^2$$

and $\widehat{\tau}_a = 0$.

Various general algorithms for solving the maximum likelihood and the REML estimating equations are reviewed by Harville (1977).

The method of Henderson and Harville. Another approach developed by Henderson (1963, 1973) and Harville (1977) is based on obtaining best linear unbiased estimators of μ and a for known σ_a and σ_e and then using these estimators to iteratively update the variance estimators. For known $\sigma_a = \widetilde{\sigma}_a$ and $\sigma_e = \widetilde{\sigma}_e$, the best linear unbiased estimators of μ and a are obtained by solving the penalised least squares problem of minimising

$$(y - \mu 1_{mg} - Za)^T (y - \mu 1_{mg} - Za)/\widetilde{\sigma}_e^2 + a^T a/\widetilde{\sigma}_a^2 \, .$$

Here, large random effects are penalised. Note that this problem is equivalent to formally extending model (1.5) to include g additional pseudo-observations of zero value and treating the effects as if they were fixed parameters,

$$\begin{pmatrix} y \\ 0 \end{pmatrix} = \begin{pmatrix} \mu 1_{mg} \\ 0 \end{pmatrix} + \begin{pmatrix} Z \\ I_g \end{pmatrix} a + \begin{pmatrix} e \\ d \end{pmatrix}.$$

Here, $e \sim N_{mg}(0, \widetilde{\sigma}_e^2 I_{mg})$ as before, and $d \sim N_g(0, \widetilde{\sigma}_a^2 I_g)$, independently. (A somewhat confusing aspect of this formal fixed-effects model is that necessarily $\widehat{d} = -\widehat{a}$.)

The normal equations are easily shown to be

$$\begin{aligned}\sum_{i=1}^{g}\sum_{j=1}^{m}(y_{ij}-\widehat{\mu}-\widehat{a}_i)/\widetilde{\sigma}_e^2 &= 0\\ \sum_{j=1}^{m}(y_{ij}-\widehat{\mu}-\widehat{a}_i)/\widetilde{\sigma}_e^2-\widehat{a}_i/\widetilde{\sigma}_a^2 &= 0\ , \quad i=1,\ldots,g\ ,\end{aligned} \tag{2.12}$$

or, on rearranging terms,

$$\begin{pmatrix} \dfrac{mg}{\widetilde{\sigma}_e^2} & \dfrac{m}{\widetilde{\sigma}_e^2}1_g^T \\ \dfrac{m}{\widetilde{\sigma}_e^2}1_g & \left(\dfrac{1}{\widetilde{\sigma}_a^2}+\dfrac{m}{\widetilde{\sigma}_e^2}\right)I_g \end{pmatrix}\begin{pmatrix}\widehat{\mu}\\ \widehat{a}\end{pmatrix} = \frac{m}{\widetilde{\sigma}_e^2}\begin{pmatrix} g\bar{y}_{\cdot\cdot}\\ \bar{y}_{1\cdot}\\ \vdots \\ \bar{y}_{g\cdot}\end{pmatrix}\ . \tag{2.13}$$

The solution to these equations is

$$\begin{aligned}\widehat{\mu} &= \bar{y}_{\cdot\cdot}\\ \widehat{a}_i &= \frac{\widetilde{\sigma}_a^2}{\widetilde{\sigma}_a^2+\widetilde{\sigma}_e^2/m}\ (\overline{y}_{i\cdot}-\bar{y}_{\cdot\cdot})\ , \quad i=1,\ldots,g\ .\end{aligned} \tag{2.14}$$

Henderson and Harville propose updating the estimates of σ_a^2 and σ_e^2 by

$$\widehat{\sigma}_a^2 = \frac{\widetilde{\sigma}_a^2+\widetilde{\sigma}_e^2/m}{\widetilde{\sigma}_a^2}\ \frac{1}{g}\ \widehat{a}^T\widehat{a} = \frac{\widetilde{\sigma}_a^2}{\widetilde{\sigma}_a^2+\widetilde{\sigma}_e^2/m}\ \frac{1}{g}\sum_{i=1}^{g}(\overline{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2 \tag{2.15}$$

$$\begin{aligned}\widehat{\sigma}_e^2 &= \frac{1}{mg}\ y^T(y-1_{mg}\widehat{\mu}-Z\widehat{a})\\ &= \frac{1}{mg}\left[\sum_{i=1}^{g}\sum_{j=1}^{m}(y_{ij}-\overline{y}_{i\cdot})^2+\frac{\widetilde{\sigma}_a^2}{\widetilde{\sigma}_a^2+\widetilde{\sigma}_e^2/m}\ \frac{1}{g}\sum_{i=1}^{g}(\overline{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2\right]\ .\end{aligned} \tag{2.16}$$

It is straightforward to show that the fixed points of equations (2.15) and (2.16) are the maximum likelihood estimators of σ_a^2 and σ_e^2. Indeed, the Henderson-Harville approach can be viewed as a general algorithm for solving the maximum likelihood equations (2.6), (2.7) and (2.8).

The REML approach corresponds to replacing (2.15, 2.16) by

$$\widehat{\sigma}_a^2 = \frac{\widetilde{\sigma}_a^2+\widetilde{\sigma}_e^2/m}{\widetilde{\sigma}_a^2}\ \frac{1}{g-1}\ \widehat{a}^T\widehat{a} \tag{2.17}$$

$$\widehat{\sigma}_e^2 = \frac{1}{mg-1}\ y^T(y-1_{mg}\widehat{\mu}-Z\widehat{a})\ , \tag{2.18}$$

see Harville (1977).

One advantage of the Henderson-Harville approach is that it yields estimates of the random effects a explicitly. The variance estimates are non-negative provided the initial estimates are also non-negative.

3 Robust Proposals

Robust variance component estimators can be constructed by modifying one of the classical estimation procedures. Although the classical procedures lead to similar or identical estimators in the present simple problem, it is apparent that they provide different points of departure for robustification and that robust procedures derived from them need not be identical.

Rocke's procedure. Rocke (1983) proposed a modification to the analysis of variance estimator (2.3, 2.4). He proposed that we obtain M-estimates $\widehat{\mu+a}_1, \ldots, \widehat{\mu+a}_g$ by solving the system of equations

$$\sum_{j=1}^{m} \psi_e\left(\frac{y_{ij} - (\widehat{\mu+a_i})}{\tilde{\sigma}_e}\right) = 0 \;, \quad i = 1, \ldots, g \;, \tag{3.1}$$

where $\tilde{\sigma}_e$ is an initial estimate of σ_e, and ψ_e is a bounded antisymmetric function, e. g., the Huber function $\psi_b(r) = r \ \min(1, b/|r|)$ with a suitable tuning constant b. Then obtain $\hat{\mu}$ by solving

$$\sum_{i=1}^{g} \psi_a\left(\frac{\widehat{\mu+a}_i - \hat{\mu}}{\tilde{\sigma}_m}\right) = 0 \;,$$

where $\tilde{\sigma}_m$ is a scale estimate based on $\widehat{\mu+a}_1, \ldots, \widehat{\mu+a}_g$ and ψ_a is chosen equal to ψ_e by Rocke. Let $\hat{a}_i = \widehat{\mu+a}_i - \hat{\mu}$. Next define pseudo-observations

$$\hat{y}_{ij} = \hat{\mu} + \sqrt{K_a}\tilde{\sigma}_m \psi_a\left(\frac{\hat{a}_i}{\tilde{\sigma}_m}\right) + \sqrt{K_e}\tilde{\sigma}_e \psi_e\left(\frac{y_{ij} - (\widehat{\mu+a_i})}{\tilde{\sigma}_e}\right) \;,$$

where $\sqrt{K_e}$ and $\sqrt{K_a}$ are suitable constants. (Rocke used the constants given by Huber (1981, p.174) who suggested that they be used to reduce bias in the estimation of asymptotic covariance of a regression M-estimator; the motivation for their use is unclear in our present context.) Finally, σ_a^2

and σ_e^2 are estimated by applying the analysis of variance estimator (2.3, 2.4) to the pseudo-observations $\widehat{y}_{ij}$ instead of the observations y_{ij}. This results in

$$\begin{aligned}
\widehat{\sigma}_e^2 &= \widetilde{\sigma}_e^2 K_e \frac{1}{g(m-1)} \sum_{i=1}^{g} \sum_{j=1}^{m} \psi_e \left(\frac{y_{ij} - (\widehat{\mu + a_i})}{\widetilde{\sigma}_e} \right)^2 \\
\widehat{\sigma}_a^2 &= \widetilde{\sigma}_m^2 K_a \frac{1}{g-1} \sum_{i=1}^{g} \psi_a \left(\frac{\widehat{a}_i}{\widetilde{\sigma}_m} \right)^2 - \widehat{\sigma}_e^2/m \,. \qquad (3.2)
\end{aligned}$$

Modifications of Rocke's method. This procedure could be iterated using

$$\widehat{\sigma}_m^2 = \widehat{\sigma}_a^2 + \widehat{\sigma}_e^2/m$$

for $\widetilde{\sigma}_m^2$ rather than re-calculating it from the $\widehat{\mu + a_i}$'s. The fixed point of such an iteration fulfills

$$\begin{aligned}
\sum_{j=1}^{m} \psi_e \left(\frac{y_{ij} - \widehat{\mu} - \widehat{a}_i}{\widehat{\sigma}_e} \right) &= 0 \\
\sum_{i=1}^{g} \psi_a \left(\frac{\widehat{a}_i}{\widehat{\sigma}_m} \right) &= 0 \\
K_a \, \frac{1}{g-1} \sum_{i=1}^{g} \psi_a \left(\frac{\widehat{a}_i}{\widehat{\sigma}_m} \right)^2 &= 1 \qquad (3.3) \\
K_e \, \frac{1}{g(m-1)} \sum_{i=1}^{g} \sum_{j=1}^{m} \psi_e \left(\frac{y_{ij} - \widehat{\mu} - \widehat{a}_i}{\widehat{\sigma}_e} \right)^2 &= 1 \,. \qquad (3.4)
\end{aligned}$$

The parameters we want to estimate efficiently are σ_a^2 and σ_e^2, whereas μ and a_i play the role of nuisance "parameters". General experience suggests that the latter should be estimated as robustly as possible, which may lead to $\psi_a(r) = \psi_e(r) = \mathrm{sign}(r)$. Since this $\psi-$function cannot be used in (3.3, 3.4), we suggest generalising them to

$$\sum_{i=1}^{g} \chi_a \left(\frac{\widehat{a}_i}{\widehat{\sigma}_m} \right) = 0 \qquad \sum_{i=1}^{g} \sum_{j=1}^{m} \chi_e \left(\frac{y_{ij} - \widehat{\mu} - \widehat{a}_i}{\widehat{\sigma}_e} \right) = 0 \,, \qquad (3.5)$$

where χ_a and χ_e are symmetric functions, monotone non-decreasing for positive arguments, and may depend on m and g. For example, $\chi_a(r) =$

$\psi_a(r)^2 - (g-1)/(gK_a)$, yields (3.2) again, but the tuning constant for χ_a can be larger than the constant used for ψ_a – and similarly for χ_e.

A "most robust" choice in this class leads to

$$\begin{aligned}\widehat{\mu{+}a_i} &= \text{med}_j(y_{ij}) \qquad \widehat{\mu} = \text{med}_i(\widehat{\mu{+}a_i}) \\ \widehat{\sigma}_e &= \text{med}_{ij}(|y_{ij} - (\widehat{\mu{+}a_i})|)\,/\,0.6745 \\ \widehat{\sigma}_a^2 &= \max\left\{\,[\text{med}_i(|\widehat{a}_i|)/0.6745]^2 - \widehat{\sigma}_e^2/m\;,\;\varepsilon\widehat{\sigma}_e^2\,\right\}\end{aligned} \tag{3.6}$$

with $\varepsilon = 0$. This estimator appears suitable as a starting point for iterative algorithms if the minimum ε for $\widehat{\sigma}_a^2$ is set to a positive value.

If these estimators are to be Fisher-consistent as $g \to \infty$ under the normal model, the constants K_a and K_e must be chosen appropriately, or the χ–functions must fulfill the asymptotic versions of (3.5),

$$\int \chi_a\left(\widehat{a}(y_i)/\sigma_a\right) dF(y_i;\sigma_a) = 0 \qquad \int \chi_e(y_{i1} - \widehat{a}(y_i))^2 dF(y_i;\sigma_a) = 0\;.$$

where $F(y_i;\sigma_a)$ is the model distribution with parameters σ_a, $\sigma_e = 1$, and $\mu = 0$, $s_a^2 = \sigma_a^2 + \sigma_e^2/m$, and $\widehat{a}(y_i)$ is the solution to (3.1) (with $\sigma_e = 1$, $\mu = 0$). Note that $\widehat{a}(y_i) - a_i = \widehat{a}(e_i)$ is independent of a_i and is distributed as the location M-estimator given by ψ_e, obtained for m independent standard normal variables. Similarly, the distribution of $y_{i1} - \widehat{a}(y_i)$ is independent of a_i and is the distribution of a "residual" from the location M-estimator. Thus, the two equations can be rewritten as

$$\int h(z\sigma_a)\; d\Phi(z) = 0$$

$$\int \chi_e(z_1 - \widehat{a}(z_1,\ldots,z_m))\; d\Phi_m(z_1,\ldots,z_m) = 0\;,$$

where

$$h(a) = \int \chi_a\left(\frac{a + \widehat{a}(z_1,\ldots,z_m)}{\sigma_a}\right) d\Phi_m(z_1,\ldots,z_m)\;.$$

Since these equations involve m-dimensional integration, it may be worthwile to find approximate solutions for moderate to large m. If m is large enough for neglecting $\widehat{a}(z_1,\ldots,z_m)$ in the two integrals, then the usual requirement $\int \chi(r) d\Phi(r) = 0$ results for both χ_a and χ_e.

Fellner's method. Two possible modifications to the preceding proposal are to improve the estimation of $a_1, \ldots, a_g$ and to apply alternative estimators of σ_a^2 and σ_e^2 to the resulting estimates. Fellner (1986) proposed doing this by modifying the Henderson–Harville approach: the a's are estimated by modifying the best linear unbiased estimator and then are used in modified REML estimators to estimate the variance components. Formally, given values of $\widetilde{\mu}$, $\widetilde{a}$, $\widetilde{\sigma}_a^2$ and $\widetilde{\sigma}_e^2$, compute the pseudo-observations

$$\widehat{y}_{ij} = \widetilde{\mu} + \widetilde{a}_i + \widetilde{\sigma}_e \psi_e \left(\frac{y_{ij} - \widetilde{\mu} - \widetilde{a}_i}{\widetilde{\sigma}_e} \right) \tag{3.7}$$

$$\widehat{O}_i = \widetilde{a}_i - \widetilde{\sigma}_a \psi_a \left(\frac{\widetilde{a}_i}{\widetilde{\sigma}_a} \right) . \tag{3.8}$$

(Fellner chooses ψ_e and ψ_a equal to a single Huber function.) Then update the estimators of μ and a by solving (2.13) for these pseudo-observations,

$$\begin{pmatrix} \dfrac{mg}{\widetilde{\sigma}_e^2} & \dfrac{m}{\widetilde{\sigma}_e^2} 1_g^T \\ \dfrac{m}{\widetilde{\sigma}_e^2} 1_g & \left(\dfrac{1}{\widetilde{\sigma}_a^2} + \dfrac{m}{\widetilde{\sigma}_e^2} \right) I_g \end{pmatrix} \begin{pmatrix} \widehat{\mu} \\ \widehat{a} \end{pmatrix} = \begin{pmatrix} \dfrac{m}{\widetilde{\sigma}_e^2} g \bar{\hat{y}}_{\cdot\cdot} \\ \dfrac{m}{\widetilde{\sigma}_e^2} \bar{\hat{y}}_{1\cdot} + \dfrac{1}{\widetilde{\sigma}_a^2} \widehat{O}_1 \\ \vdots \\ \dfrac{m}{\widetilde{\sigma}_e^2} \bar{\hat{y}}_{g\cdot} + \dfrac{1}{\widetilde{\sigma}_a^2} \widehat{O}_g \end{pmatrix} . \tag{3.9}$$

The equation corresponding to (2.14) reads

$$\begin{aligned} \widehat{\mu} &= \bar{\hat{y}}_{\cdot\cdot} - \bar{\hat{O}}. \\ \widehat{a}_i &= \frac{\widetilde{\sigma}_a^2}{\widetilde{\sigma}_a^2 + \widetilde{\sigma}_e^2/m} \left(\bar{\hat{y}}_{i\cdot} - \bar{\hat{y}}_{\cdot\cdot} - [\hat{O}_i - \bar{\hat{O}}_{\cdot}] \right) + \hat{O}_i \ , \quad i = 1, \ldots, g \ . \end{aligned}$$

Using these new values for μ and a, update the estimates of σ_a^2 and σ_e^2 by

$$\widehat{\sigma}_a^2 = (\widetilde{\sigma}_a^2 + \widetilde{\sigma}_e^2/m) K_a \frac{1}{(g-1)} \sum_{i=1}^{g} \psi_a \left(\frac{\widehat{a}_i}{\widetilde{\sigma}_a} \right)^2 \tag{3.10}$$

$$\widehat{\sigma}_e^2 = \widetilde{\sigma}_e^2 K_e \frac{1}{mg-1} \sum_{i=1}^{g} \sum_{j=1}^{m} \psi_e \left(\frac{y_{ij} - \widehat{\mu} - \widehat{a}_i}{\widetilde{\sigma}_e} \right)^2 , \tag{3.12}$$

where

$$K_a^{-1} = \int \psi_a(x)^2 \, d\Phi(x) \ , \qquad K_e^{-1} = \int \psi_e(x)^2 \, d\Phi(x) \ , \tag{3.11}$$

and Φ is the standard normal distribution. The procedure is then iterated to convergence.

Let us examine the fixed point of this algorithm. It is easy to derive from (3.7, 3.8, 3.9) with $\widehat{a} = \widetilde{a}$ and $\widehat{\mu} = \widetilde{\mu}$ that

$$\sum_{i=1}^{g}\sum_{j=1}^{m}\psi_e\left(\frac{y_{ij}-\widehat{\mu}-\widehat{a}_i}{\widehat{\sigma}_e}\right) = 0 \tag{3.13}$$

$$\frac{1}{\widehat{\sigma}_e}\sum_{j=1}^{m}\psi_e\left(\frac{y_{ij}-\widehat{\mu}-\widehat{a}_i}{\widehat{\sigma}_e}\right) - \frac{1}{\widehat{\sigma}_a}\psi_a\left(\frac{\widehat{a}_i}{\widehat{\sigma}_a}\right) = 0, \quad i = 1, ..., g\ , \tag{3.14}$$

which is a direct "robustification" of (2.12). In (3.10) and (3.12), the fixed point condition leads to

$$\frac{\widehat{\sigma}_a^2 + \widehat{\sigma}_e^2/m}{\widehat{\sigma}_a^2}K_a\frac{1}{g-1}\sum_{i=1}^{g}\psi_a\left(\frac{\widehat{a}_i}{\widehat{\sigma}_a}\right)^2 = 1 \tag{3.15}$$

$$K_e\frac{1}{mg-1}\sum_{i=1}^{g}\sum_{j=1}^{m}\psi_e\left(\frac{y_{ij}-\widehat{\mu}-\widehat{a}_i}{\widehat{\sigma}_e}\right)^2 = 1\ . \tag{3.16}$$

Modifications of Fellner's estimator. Equation (3.15) corresponds to (2.17), but (3.16) is different from (2.18) for $\psi(r) = r$. A streamlined modification can be obtained by writing (2.18) as

$$\widehat{\sigma}_e^2 = \frac{1}{mg-1}\left[\sum_{i=1}^{g}\sum_{j=1}^{m}(y_{ij}-\widehat{\mu}-\widehat{a}_i)^2 + \frac{\widetilde{\sigma}_e^2}{\widetilde{\sigma}_a^2}\sum_{i=1}^{g}\widehat{a}_i^2\right]$$

(obtained by straightforward calculation) and therefore replacing (3.12) by

$$\widehat{\sigma}_e^2 = \frac{\widetilde{\sigma}_e^2}{mg-1}\left[K_e\sum_{i=1}^{g}\sum_{j=1}^{m}\psi_e\left(\frac{y_{ij}-\widehat{\mu}-\widehat{a}_i}{\widetilde{\sigma}_e}\right)^2 + K_a\sum_{i=1}^{g}\psi_a\left(\frac{\widehat{a}_i}{\widetilde{\sigma}_a}\right)^2\right]\ . \tag{3.17}$$

The corresponding fixed-point condition can be written, using (3.15), as

$$\frac{1}{mg-1}\left[K_e\sum_{i=1}^{g}\sum_{j=1}^{m}\psi_e\left(\frac{y_{ij}-\widehat{\mu}-\widehat{a}_i}{\widehat{\sigma}_e}\right)^2 + (g-1)\frac{\widehat{\sigma}_a^2}{\widehat{\sigma}_a^2+\widehat{\sigma}_e^2/m}\right] = 1\ . \tag{3.18}$$

Fisher consistency requires that asymptotic versions of (3.15, 3.18) hold, leading to

$$K_a^{-1} = \frac{\sigma_a^2 + 1/m}{\sigma_a^2} \int \psi_a \left(\frac{\widehat{a}(y_i)}{\sigma_a} \right)^2 dF(y_i; \sigma_a)$$

$$K_e^{-1} = \frac{m\sigma_a^2 + 1}{(m-1)\sigma_a^2 + 1} \int \psi_e(y_{i1} - \widehat{a}(y_i))^2 dF(y_i; \sigma_a) \ .$$

where $F(y_i; \sigma_a)$ is again the model distribution with parameters σ_a, $\sigma_e = 1$, and $\mu = 0$, and $\widehat{a}(y_i)$ is the solution to

$$\sum_{j=1}^{m} \psi_e(y_{ij} - \widehat{a}(y_i)) - \frac{1}{\sigma_a} \psi_a(\widehat{a}(y_i)/\sigma_a) = 0 \ .$$

In this case, the distribution of $\widehat{a}(y_i) - a_i$ depends on a_i, and no essential simplification appears possible. The "constants" K_a, K_e thus depend on σ_a/σ_e and on m. This makes the estimation procedure more difficult. It also implies that Fellner's estimator and the modification just described are inconsistent if K_a and K_e are fixed constants. The bias vanishes for $m \to \infty$ if (3.11) is used.

Upon closer examination, there is a slightly different and more convincing way to robustify the Henderson-Harville approach. Equation (3.15) suggests and calculation shows that the variance of the estimated effects $\widehat{a}_i$ in the normal case with $\psi_a(r) = r$ is not σ_a but

$$\sigma_a^{*2} = \sigma_a^2 \frac{g-1}{g[1 + \sigma_e^2/(m\sigma_a^2)]} \ . \tag{3.19}$$

In fact, (2.17) can then be written as $\sum_{i=1}^{g} [(\widehat{a}_i/\widehat{\sigma}_a^*)^2 - 1] = 0$. A straightforward robustification is

$$\sum_{i=1}^{g} \chi_a \left(\frac{\widehat{a}_i}{\widehat{\sigma}_a^*} \right) = 0 \ . \tag{3.20}$$

Equation (3.18) with normal data and $\psi_e(r) = r$ leads to

$$\mathrm{var}(y_{ij} - \widehat{\mu} - \widehat{a}_i) = \sigma_e^{*2} = \sigma_e^2 \left(1 - \frac{1}{mg} - \frac{g-1}{mg} \frac{1}{1 + \sigma_e^2/(m\sigma_a^2)} \right) \ , \tag{3.21}$$

and $\sum_{i=1}^{g} \sum_{j=1}^{m} ([y_{ij} - \widehat{\mu} - \widehat{a}_i/\widehat{\sigma}_e^*]^2 - 1) = 0$. The robust version is

$$\sum_{i=1}^{g} \sum_{j=1}^{m} \chi_e \left(\frac{y_{ij} - \widehat{\mu} - \widehat{a}_i}{\widehat{\sigma}_e^*} \right) = 0 \ . \tag{3.22}$$

The conditions for Fisher-consistency are

$$\int \chi_a\left(\frac{\widehat{a}(y_i)}{\sigma_a}\sqrt{1+1/(m\sigma_a^2)}\right)dF(y_i;\sigma_a)=0$$

$$\int \chi_e\left([y_{i1}-\widehat{a}(y_i)]\sqrt{\frac{1+m\sigma_a^2}{1+(m-1)\sigma_a^2}}\right)dF(y_i;\sigma_a)=0\ .$$

For large m, this simplifies to

$$\int \chi_a(a)d\Phi(a)=0\ ,\qquad \int \chi_e(e)d\Phi(e)=0\ . \tag{3.23}$$

Heuristic arguments imply that the asymptotic bias ($g\to\infty$) when using (3.23) for finite m should be smaller in the case of this second modification than for Fellner's original proposal and our first modification. (If $\chi_e(r)=K_e\psi_e(r)^2-1$ and similarly for χ_a, then (3.23) is equivalent to (3.11).)

Besides their simplicity, these equations lead to a more direct interpretability of the χ functions since their arguments are standard normal variates in the case of the central model. This suggests that σ_a and σ_e are also replaced by σ_a^* and σ_e^* in (3.13) and (3.14). These two modified equations, together with (3.20) and (3.22), determine estimates $\widehat{\mu}$, $\widehat{a},\widehat{\sigma}_a^*$, and $\widehat{\sigma}_e^*$. The estimates for σ_a and σ_e are then found by solving (3.19) and (3.21). An iterative algorithm for doing this is based on replacing current values $\widetilde{\sigma}_a^2$, $\widetilde{\sigma}_e^2$ by

$$\widehat{\sigma}_a^2=\widehat{\sigma}_a^{*2}\left(1+\widetilde{\sigma}_e^2/(m\widetilde{\sigma}_a^2)\right)\ g/(g-1)$$

$$\widehat{\sigma}_e^2=\widehat{\sigma}_e^{*2}\Big/\left(1-\frac{1}{mg}-\frac{g-1}{mg}\frac{1}{1+\widetilde{\sigma}_e^2/(m\widetilde{\sigma}_a^2)}\right)\ .$$

Optimal B-robust estimator. The general procedure called optimal B-robust estimation can be readily applied to the multivariate formulation. Let us treat $\tau_e = \sigma_e^2$, $\tau_m = \sigma_a^2 + \sigma_e^2/m$ and μ separately in the sense of a "partitioned parameter" (Hampel et al., 1986). It is useful to standardize the likelihood scores (2.9) by the square root of the Fisher infromation (2.11) before applying Huber functions. This leads to

$$\begin{aligned} \sum_{i=1}^{g} \psi_{b_m}\left(\frac{(\overline{y}_{i\cdot} - \mu)^2/\tau_m - 1}{\sqrt{2}} - \alpha_m\right) &= 0 \\ \sum_{i=1}^{g} \psi_{b_e}\left(\frac{(y_i - 1_m\overline{y}_{i\cdot})^T(y_i - 1_m\overline{y}_{i\cdot})/\tau_e - (m-1)}{\sqrt{2(m-1)}} - \alpha_e\right) &= 0 \\ \sum_{i=1}^{g} \psi_{b_\mu}\left(\frac{\overline{y}_{i\cdot} - \mu}{\sqrt{\tau_m}}\right) &= 0\ , \end{aligned} \tag{3.24}$$

where α_e and α_m are determined to obtain Fisher consistency at the (normal) model,

$$\int \psi_{b_m}(\frac{u-1}{\sqrt{2}} - \alpha_m)\ d\chi_1^2(u) = 0 \qquad \int \psi_{b_e}(\frac{u-(m-1)}{\sqrt{2(m-1)}} - \alpha_e)\ d\chi_{m-1}^2(u) = 0\ .$$

The tuning constants b_m, b_e, and b_μ can be chosen arbitrarily (but positive). They determine the robustness and efficiency of the estimator.

The simplicity of this result relies on the anti-symmetry of $s^{(\mu)}$ and the independence of $s^{(\tau_m)}$ and $s^{(\tau_e)}$, which leads to a diagonal matrix A in (4.4.1) of Hampel et al. (1986).

The most B-robust estimator is obtained by letting the b's go to zero (after dividing the equations by them). This leads to

$$\begin{aligned} \widehat{\mu} &= \operatorname{med}(\bar{y}_{1\cdot}, \ldots, \bar{y}_{g\cdot}) \\ \widehat{\tau}_m &= \operatorname{med}(|\bar{y}_{1\cdot} - \widehat{\mu}|^2, \ldots, |\bar{y}_{g\cdot} - \widehat{\mu}|^2)/0.6745^2 \\ \widehat{\tau}_e &= \operatorname{med}(\textstyle\sum_{j=1}^m (y_{1j} - \bar{y}_{1\cdot})^2, \ldots, \sum_{j=1}^m (y_{gj} - \bar{y}_{g\cdot})^2)/\operatorname{med}(\chi_{m-1}^2)\ . \end{aligned} \tag{3.25}$$

If the influence function is not restricted in this componentwise manner, but its Euclidean norm is bounded instead, or if another parametrisation is used, the result looks considerably more complicated (since A is no longer diagonal). These cases are therefore not pursued here.

A somewhat surprising feature of the above results is the appearance of group means and within-group sums of squares in the formulas. This is a consequence of treating the vector of all y_{ij}'s in a group i as one observation and of using the normal model as the "central model". A simple remedy is to replace means and sums of squares by robust versions. This leads to proposals like Rocke's.

Maximum likelihood for long-tailed distributions. Assume now that a_i and e_{ij} have arbitrary distributions with densities written as

$$\begin{aligned} f_a(a;\sigma_a) &= c_a\sigma_a^{-1}\exp\{-\rho_a(a/\sigma_a)\} \\ f_e(e;\sigma_e) &= c_e\sigma_e^{-1}\exp\{-\rho_e(e/\sigma_e)\}\ . \end{aligned} \tag{3.26}$$

Then, the likelihood for an observation y_i gets

$$f(y_i;\sigma_a,\sigma_e,\mu) = \int \tilde{f}(y_i,a_i;\sigma_a,\sigma_e,\mu)da_i \tag{3.27}$$

with

$$\begin{aligned} \tilde{f}(y_i,a_i;\sigma_a,\sigma_e,\mu) =& \sigma_e^{-m}\left[\prod_{j=1}^{m}\exp\left\{-\rho_e\left(\frac{y_{ij}-\mu-a_i}{\sigma_e}\right)\right\}\right] \\ & \cdot\sigma_a^{-1}\exp\left\{-\rho_a\left(\frac{a_i}{\sigma_a}\right)\right\}\cdot\text{const}\ . \end{aligned} \tag{3.28}$$

For straightforward maximum likelihood estimation, (3.27) can be maximised numerically with respect to σ_a, σ_e, and μ. The likelihood scores are

$$\begin{aligned} s^{(\sigma_a)}(y_i;\sigma_a,\sigma_e,\mu) &= \sigma_a^{-1}\left[E^*\zeta_a\left(\frac{*}{\sigma_a}\right)-1\right] \\ s^{(\sigma_e)}(y_i;\sigma_a,\sigma_e,\mu) &= \sigma_e^{-1}\left[\sum_{j=1}^{m}E^*\zeta_e\left(\frac{y_{ij}-\mu-*}{\sigma_e}\right)-m\right] \\ s^{(\mu)}(y_i;\sigma_a,\sigma_e,\mu) &= \sigma_e^{-1}\sum_{j=1}^{m}E^*\rho_e'\left(\frac{y_{ij}-\mu-*}{\sigma_e}\right)\ , \end{aligned}$$

where $\zeta(r) = \rho'(r)r$ and

$$E^*h(*) = \int h(a)\tilde{f}(y_i,a;\sigma_a,\sigma_e,\mu)\ da\Big/ f(y_i;\sigma_a,\sigma_e,\mu)$$

is the expectation of h over the posterior density of a_i, given the observations y_{ij} of the i-th group.

In the normal case, the integration in (3.27) can be carried out explicitly to arrive at (2.5). The Henderson-Harville method is based on the fact that integration can be replaced by maximisation. This can be seen by noting that (a_i, y_i) is multivariate normal and therefore the integrand $\widetilde{f}(y_i, a_i; \sigma_a, \sigma_e, \mu)$, viewed as a function of a_i, is a scaled normal density with variance σ_e^2/τ_m, independently of y_i. Thus, the integral is proportional to the maximum, which is achieved at $\widehat{a}_i = (\overline{y}_{i\cdot} - \mu)\sigma_a^2/\tau_m$, see (2.14). The proportionality constant is independent of μ, and therefore the maximum of $\widetilde{f}(y_i, \widehat{a}_i; \sigma_a, \sigma_e, \mu)$ is assumed at the same μ as the maximum of $f(y_i; \sigma_a, \sigma_e, \mu)$.

In the non-normal case, the two operations – integration and maximisation – lead to different procedures. In order to simplify matters, we may replace integration by maximisation anyway. This amounts to maximising (3.28) with respect to a_i and μ for fixed σ_a and σ_e, which leads to (3.13, 3.14). If we further assume that the factor converting the maximal $\widetilde{f}$ into the integral f behaves similarly to the normal case, equations similar to (3.15, 3.16) can be obtained, but with ζ instead of $\psi^2 = \rho'^2$. It would be interesting to study the actual shapes of posterior distributions in the non-normal case.

This maximum likelihood procedure estimates the parameters σ_a ans σ_e defined in (3.26). It will turn out to be robust if f_a and f_e are long-tailed. If it is used as a robust estimator for the parameters of the normal model, some consistency correction is essential.

4 A Small Simulation Study

We implemented a small simulation study to investigate the performance of some of the readily computable estimators presented above. All the data was generated using the model (1.1) with $m = 4$ and $g = 5$ or 20 producing overall sample sizes of $mg = 20$ and 80. When $g = 5$, there are only $g - 1 = 4$ degrees of freedom for estimating σ_a^2 and $g(m-1) = 15$ for estimating σ_e^2 while with $g = 20$, there are 19 degrees of freedom for estimating σ_a^2 and 60 for estimating σ_e^2. Thus the effective sample sizes for estimating σ_a^2 are small and moderate in the two cases. We allowed for contamination in either or both of the random effect distributions. These were taken to be of the form

$$(1 - \varepsilon)N(0, 2^2) + \varepsilon N(0, 8^2)$$

with $\varepsilon = 0$ or 0.1, producing four possible combinations (0/0), (0/0.1), (0.1/0) and (0.1/0.1) which correspond to uncontaminated data, contamination in the second component (the e distribution), contamination in the first component (the a distribution), and contamination in both components. The variance of the distribution is 4 when $\varepsilon = 0$ and 10 when $\varepsilon = 0.1$, the effect of the contamination being to increase the variance. However, in the presence of contamination, our interest is in estimating the underlying variance which is 4. The parameters were chosen in an attempt to ensure that most estimates of σ_a^2 are off the boundary $\sigma_a^2 = 0$ and that the contaminated variance is far enough from the uncontaminated variance to make clear distinctions between the estimators possible. Each situation was replicated 500 times.

Eleven estimators of the pair (σ_a^2, σ_e^2) were included in the study. These can be arranged in four groups:

- Classical:
  - ANOVA, the truncated ANOVA estimator (2.3)-(2.4),
  - REML, the reduced maximum likelihood estimator described in Section 2. It equals the ANOVA estimator except when the estimate of σ_a^2 is on the boundary (in which case the estimate of σ_e^2 changes).
  - MLE, the maximum likelihood estimator, (2.10). It produces the same estimate of σ_e^2 as the ANOVA and REML estimator except when the estimate σ_a^2 is on the boundary. The estimate of σ_a^2 is slightly different (biased).
- Fellner's type:
  - Fellner's proposal,
  - Fellner.mod, the modification given in (3.17),
  - Fellner.new, the estimator given by (3.19) - (3.22) and (3.7, 3.8) replacing $\tilde{\sigma}_a$, $\tilde{\sigma}_e$ by $\tilde{\sigma}_a^*$, $\tilde{\sigma}_e^*$;
- Rocke's type:
  - Rocke's proposal (3.2),
  - Rocke.it, the iterated version given by (3.3) - (3.4),
  - Rocke.most, the limiting most roust case (3.6);

- B-optimal estimators:
  - Bopt.ave, the B-optimal estimator given by (3.24),
  - Bmost.ave, the limiting most robust version (3.25),
  - Bopt.med, obtained by replacing means $\bar{y}$ by medians in (3.24),
  - Bmost.med, doing the same in (3.25).

All the robust estimators other than the most robust versions had the tuning constant of Huber's ψ-function set to 1.34, the widely used value for location estimation. The consistency constants given by (3.11), which were used for all Fellner type estimators, are $K_a = K_e = 1.408$.

The simulated distributions of the estimators for both variance components turned out to be asymmetric and long tailed. Figure 1 displays a typical shape. The distributions will therefore be summarised by robust analogues of bias and mean squared error based on medians,

$$B = \text{med } (\widehat{\sigma}^2) - \sigma_0^2$$
$$\text{MD} = \text{med}(|\widehat{\sigma}^2 - \sigma_0^2|) \ ,$$

where $\sigma_0^2 = 4$ is the true value for both σ_a^2 and σ_e^2 in our study.

[Figure 1 about here]

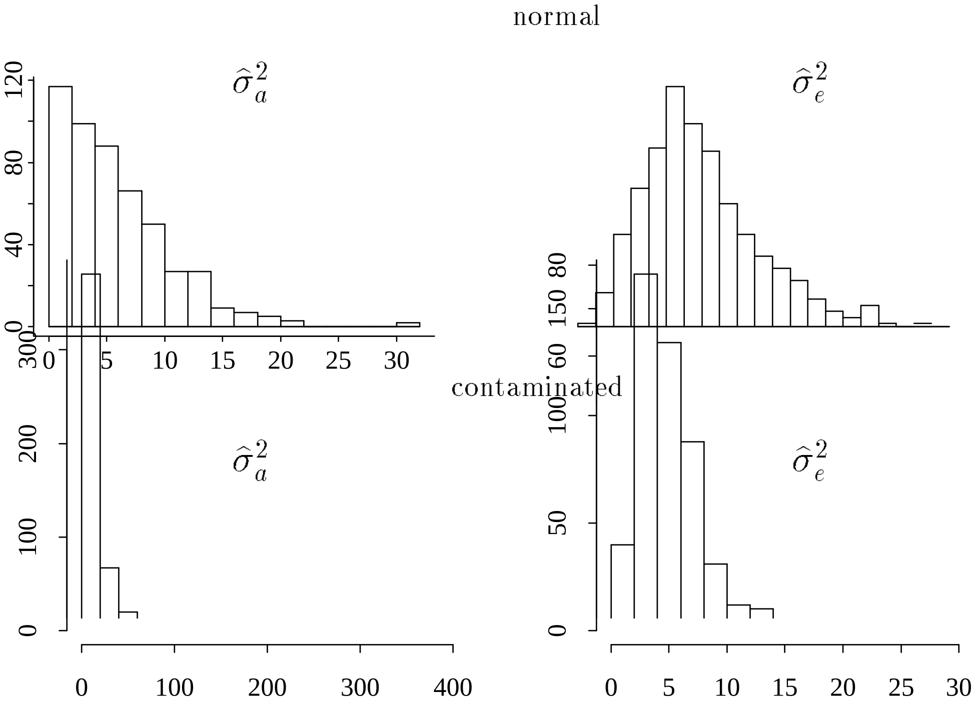
A robust efficiency measure is

$$\text{REF}_T = \text{MD}_{\text{ANOVA}} / \text{MD}_T \ ,$$

the ratio of the MD value of the truncated ANOVA estimator to the MD value for estimator T.

Table 1 lists these values for both variance components. Since the values in the table are on the square root of the usual mean squared error scale, they tend to be closer to one than those obtained using mean squared errors. The results can be interpreted more clearly by taking the bias figures (B) into account. They are listed in Table 2. Note that high entries in Table 1 and low entries in Table 2 are desirable.

[Tables 1 and 2 about here]


normal

contaminated

Figure 1: Simulated distributions of Fellner's $\hat{\sigma}_a^2$ and $\hat{\sigma}_e^2$, $g = 5$, for both component distributions normal (top) and both contaminated (bottom)

The classical methods all perform comparably. It is interesting to note here that the ANOVA estimate of σ_a^2 was frequently negative in spite of choosing parameter values to try to minimise the problem: Out of 500 replications, the number of negative estimates for $g = 5$ were 34, 143, 33 and 99 for the four cases respectively, while for $g = 20$ there were 0, 65, 0 and 11 negative estimates. Contaminating the e distribution leads to more non-admissible estimates than contaminating the a distribution – a plausible result. The three methods of estimating σ_e^2 produce different estimates only when the estimate of σ_a^2 is on the boundary; otherwise the truncated ANOVA and REML estimates of σ_a^2 are the same but differ from the maximum likelihood estimate which is slightly negatively biased. Contamination of only one component distribution increases both the bias and variability of the estimator of the variance of that distribution but has a substantially smaller impact on the estimator of the other variance. The bias dominates the variability except for the estimators of σ_a^2 when $g = 5$ where the opposite occurs due to the very small effective sample size. The results show clearly that these estimators do not do a good job of estimating the variance components when outliers are present. Note that for the non-parametric problem of estimating the variance of the contaminated distribution (which is 10) there is still considerable bias since zero bias for this problem corresponds to a bias of six in Table 1.

Fellner's original estimator and our modifications show similar results. Differences occur in the estimation of σ_e^2. The first modification appears to be more efficient in the uncontaminated cases, but less robust than the original proposal. The second modification, "Fellner.new," almost achieves the combination of the better of the efficiency and robustness results of the two competitors. Clearly, the shape and the tuning constant of the ψ-function determine these properties for each estimator. More research is needed to study the range of behaviour that is achieved by different ψ-functions.

For the estimators based on Rocke's proposal it is convenient to discuss the estimation of σ_a^2 and σ_e^2 separately. The iterated estimator of σ_e^2 shows both smaller bias and smaller variability than the one-step or the most robust estimators which tend to exhibit surprisingly comparable performances. The results for estimating σ_a^2 are less clear-cut. When $g = 5$, the one-step estimator tends to have smaller bias than the most robust estimator which in turn has smaller bias than the iterated estimator. However, when $g = 20$, the most robust estimator has smaller or comparable bias than the one-step estimator. The iterated estimator has lower variability than the one-step and most robust estimators which exhibit comparable varibility. Overall, the iterated estimator performs best when the a distribution is contaminated but loses to the other two estimators when it is not. The most robust estimator is never far behind the other two but is not the best in any of the situations we have considcered. In terms of estimating both σ_a^2 and σ_e^2, the iterated estimator offers the greatest gains, and when it loses out it does not do so too badly. These results do not speak clearly to the choice of the cut-off value of ψ.

Overall, the results show that the possible gains and losses are much greater than in simple location problems and large gains over what can be achieved using classical normal theory methods are possible. The greatest of these are achieved in this study by the iterated version of Rocke's estimator and Fellner's estimator. Our modification of Fellner's proposal have better efficiency at the model than these. Overall, they appear to have the potential to win when the ψ-function is allowed to vary. The B-optimal estimators do fairly well but not as well as these estimators.

References

Brownlee, K. A. (1953). *Industrial Experimentation.* Chemical Publishing Co.

Cochran, W. G. (1939). The Use of the Analysis of Variance in Enumeration by Sampling. J. Amer. Statist. Assoc. **34**, 492-510.

Fairfield Smith, H. (1936). A Discriminant Function for Plant Selection. Ann. Eugenics **7**, 240-260.

Fellner, W. H. (1986). Robust Estimation of Variance Components. Technometrics **28**, (No. 1) 51-60.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley, New York.

Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. J. Amer. Statist. Assoc. **72**, 320-340.

Henderson, C. R. (1950). Estimation of Genetic Parameters (abstract). Ann. Math. Statist. **21**, 309-310.

Henderson, C. R. (1953). Estimation of Variance and Covariance Components. Biometrics **9**, 226-252.

Henderson, C. R. (1963). Selection Index and Expected Genetic Advance. *In:* National Research Council *Statistical Genetics and Plant Breeding.* National Academy of Sciences, Publication No. 982, 141-163.

Henderson, C. R. (1973). Maximum Likelihood Estimation of Variance Components.

Huber, P. J. (1981). *Robust Statistics.* Wiley, New York.

La Motte, L. R. (1973). Quadratic Estimation of Variance Components. Biometrics **29**, 311-330.

Martin, R. D., and Zamar, R. H. (1989). Asymptotically Min-Max Robust M-Estimates of Scale for Positive Random Variables. J. Amer. Statist. Assoc. **84**, 494-501.

Patterson, H. D., and Thompson, R. (1971). Recovery of Interblock Information When Block Sizes are Unequal. Biometrika **58**, 545-554.

Pukelsheim, P. J. (1981). On the Existence of Unbiased Non-negative Estimates of Variance and Covariance Components.

Rao, C. R. (1947). General Methods of Analysis for Incomplete Block Designs. J. Amer. Statist. Assoc. **42**, 541-561.

Rao, C. R. (1970). Estimation of Heteroscedastic Variances in Linear Models. J. Amer. Statist. Assoc. **65**, 161-172.

Rao, C. R. (1971a). Estimation of Variance and Covariance Components. J. of Multivariate Analysis **1**, 257-275.

Rao, C. R. (1971b). Minimum Variance Quadratic Unbiased Estimation of Variance Components. J. of Multivariate Analysis **1**, 445-456.

Rao, C. R. and Kleffe, J. (1980). Estimation of Variance Components. *In:* Krishnaiah, P. R. (ed.) *Handbook of Statistics.* North-Holland, Amsterdam, 1-40.

Rao, C. R. and Kleffe, J. (1988). *Estimation of Variance Components and Applications.* North-Holland, Amsterdam.

Rieder, H. (1991). Robust Testing of Functionals. *In:* Stahel, W. and Weisberg, S. (eds.) *Directions in Robust Statistics and Diagnostics, vol. 2.* Springer, N.Y.

Rocke, D. M. (1983). Robust Statistical Analysis of Interlaboratory Studies. Biometrika **70**, 421-431.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics.* Wiley, New York.

Westfall, P. (1988). Robustness and Power of Tests for a Null Variance Ratio. Biometrika **75**, (No. 2) 207-214.

Yates, F. (1940). The Recovery of Inter-Block Information in Balanced Incomplete Block Designs. Ann. Eugenics **10**, 317-325.

Yates, F., and Zacopancy, I. (1935). The Estimation of the Efficiency of Sampling with Special Reference to Sampling for Yield in Cereal Experiments. J. Agric. Sci. **25**, 545-577.

Youden, W. J. (1975). Statistical Techniques for Collaborative Tests. *In:* Association of Analytical Chemists. *Statistical Manual of the Association of Official Analytical Chemists.* Association of Analytical Chemists, Arlington VA, v-63.

Zielinski, W. (1986). On Robust Estimation of Variance Components. Probab. Statist. **7**, (No. 2) 95-102.

	σ_a^2				σ_e^2			
Distribution a/e	0/0	0/.1	.1/0	.1/.1	0/0	0/.1	.1/0	.1/.1
	$g = 5$							
MD(ANOVA)	3.4678	5.8028	5.5093	5.9303	1.3878	20.5807	1.5158	21.9397
Estimator	Efficiencies							
REML	1.000	1.000	1.000	0.669	0.964	0.974	0.982	0.984
MLE	0.980	1.050	1.053	1.000	1.015	1.039	0.999	1.038
Fellner	0.843	1.185	0.929	1.000	0.720	8.992	0.743	8.776
Fellner.mod	0.829	1.137	0.929	1.001	0.841	6.177	0.842	5.738
Fellner.new	0.833	1.073	1.089	1.117	0.777	7.971	0.799	8.223
Rocke	0.976	1.337	1.153	1.255	0.501	8.269	0.485	9.029
Rocke.it	0.878	1.251	1.330	1.334	0.553	9.342	0.555	9.919
Rocke.most	0.755	1.133	1.080	1.070	0.470	7.520	0.457	8.249
Bopt.ave	0.927	0.979	1.091	1.000	0.930	4.162	0.983	4.623
Bmost.ave	0.710	0.978	1.022	1.000	0.750	5.743	0.792	6.409
Bopt.med	0.892	0.987	1.059	1.000	0.887	3.665	0.947	4.143
Bmost.med	0.713	0.978	0.982	1.000	0.729	5.141	0.746	5.741
	$g = 20$							
MD(ANOVA)	1.6388	3.3753	22.6675	21.8014	0.7103	32.6986	0.7429	34.2437
Estimator	Efficiencies							
REML	1.000	1.000	1.000	1.000	1.000	0.962	1.000	0.996
MLE	1.035	1.011	1.071	1.103	1.000	1.001	1.000	1.002
Fellner	0.878	1.608	5.803	5.287	0.416	31.630	0.412	29.969
Fellner.mod	0.880	1.575	6.032	5.324	0.754	9.015	0.799	9.221
Fellner.new	0.761	1.477	7.375	7.217	0.698	18.815	0.716	20.640
Rocke	0.976	1.871	7.526	6.571	0.262	20.089	0.269	19.752
Rocke.it	0.508	1.021	8.952	8.151	0.284	27.497	0.296	27.333
Rocke.most	0.687	1.328	6.978	7.248	0.251	17.814	0.255	17.422
Bopt.ave	0.893	1.058	7.409	2.997	0.915	6.321	0.932	6.619
Bmost.ave	0.637	0.974	7.498	4.466	0.771	9.957	0.728	10.466
Bopt.med	0.870	1.491	7.052	6.870	0.927	5.625	0.878	5.783
Bmost.med	0.630	1.098	7.214	6.878	0.733	8.454	0.687	9.111

Table 1: Robust Efficiencies of Estimators Relative to the Truncated ANOVA Estimator

	σ_a^2				σ_e^2			
Distribution a/e	0/0	0/.1	.1/0	.1/.1	0/0	0/.1	.1/0	.1/.1
Estimator	$g = 5$							
ANOVA	-0.75	-1.10	1.91	1.85	-0.03	13.81	-0.22	14.80
REML	-0.75	-1.10	1.91	1.85	-0.03	13.81	-0.22	14.80
MLE	-1.62	-2.61	0.53	-0.30	-0.03	13.81	-0.22	14.80
Fellner	0.69	0.59	3.45	3.18	-0.84	0.63	-1.04	0.71
Fellner.mod	0.49	0.20	3.50	2.91	0.34	2.15	0.16	2.55
Fellner.new	-1.03	-1.60	0.59	0.40	-0.33	1.09	-0.62	1.32
Rocke	-1.09	-1.27	0.18	0.04	-1.82	-1.14	-2.04	-1.11
Rocke.it	-1.88	-2.10	-0.99	-1.01	-1.63	-0.75	-1.79	-0.67
Rocke.most	-1.48	-1.83	-0.49	-0.58	-1.89	-1.29	-2.14	-1.19
Bopt.ave	-1.49	-0.88	0.05	1.57	0.15	3.32	-0.16	3.20
Bmost.ave	-2.07	-2.10	-1.09	0.22	0.12	2.20	-0.14	2.20
Bopt.med	-1.51	-2.82	0.40	-1.27	0.38	3.77	0.11	3.57
Bmost.med	-2.09	-3.10	-1.06	-1.64	0.38	2.66	0.07	2.51
	$g = 20$							
ANOVA	-0.09	-0.39	15.29	14.71	-0.06	22.06	-0.02	23.10
REML	-0.09	-0.39	15.29	14.71	-0.06	22.06	-0.02	23.10
MLE	-0.34	-0.82	14.27	13.34	-0.06	22.06	-0.02	23.10
Fellner	0.66	0.81	2.60	2.78	-1.15	0.10	-1.18	0.03
Fellner.mod	0.48	0.64	2.52	2.74	0.37	2.45	0.39	2.50
Fellner.new	-0.45	-0.66	1.22	0.94	-0.40	1.10	-0.43	1.06
Rocke	-0.14	0.15	1.63	2.06	-1.83	-0.98	-1.86	-1.09
Rocke.it	-2.17	-2.19	-1.45	-1.50	-1.69	-0.63	-1.69	-0.68
Rocke.most	0.03	0.16	1.33	1.47	-1.91	-1.18	-1.97	-1.25
Bopt.ave	-0.26	1.91	1.57	4.91	-0.07	3.49	-0.01	3.49
Bmost.ave	-0.55	1.38	0.51	2.99	-0.07	2.21	0.05	2.21
Bopt.med	-0.18	-0.65	1.82	1.08	0.15	3.92	0.26	3.99
Bmost.med	-0.49	-0.81	0.84	0.50	0.15	2.61	0.30	2.54

Table 2: Median Biases