Made available by Hasselt University Library in https://documentserver.uhasselt.be

Fast, Closed-form, and Efficient Estimators for Hierarchical Models with AR(1) Covariance and Unequal Cluster Sizes

Peer-reviewed author version

HERMANS, Lisa; NASSIRI, Vahid; MOLENBERGHS, Geert; Kenward, Michael G.; VAN DER ELST, Wim; AERTS, Marc & VERBEKE, Geert (2018) Fast, Closed-form, and Efficient Estimators for Hierarchical Models with AR(1) Covariance and Unequal Cluster Sizes. In: Communications in statistics. Simulation and computation, 47 (5),p. 1492-1505.

DOI: 10.1080/03610918.2017.1316395
Handle: http://hdl.handle.net/1942/25871

Fast, Closed-form, and Efficient Estimators for Hierarchical Models with AR(1) Covariance and Unequal Cluster Sizes

Lisa Hermans[1*] **Vahid Nassiri**[2] **Geert Molenberghs**[1,2]

Michael G. Kenward[3] **Wim Van der Elst**[4,1] **Marc Aerts**[1]

Geert Verbeke[2,1]

[1] *I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium*

[2] *I-BioStat, KU Leuven, B-3000 Leuven, Belgium*

[3] *Luton, United Kingdom*

[4] *Janssen Pharmaceutica, B-2340 Beerse, Belgium*

$*$ *Corresponding author: Martelarenlaan 42 B-3500 Hasselt; lisa.hermans@uhasselt.be*

Abstract

This paper is concerned with statistically and computationally efficient estimation in a hierarchical data setting with unequal cluster sizes and an AR(1) covariance structure. As in Hermans *et al.* (2016) for the compound-symmetry model, the pseudo-likelihood and split-sample methods of Fieuws and Verbeke (2006) and Molenberghs *et al.* (2011) are used. Maximum likelihood estimation for AR(1) requires numerical iteration when cluster sizes are unequal. A near optimal non-iterative procedure is proposed. Results show that the method is statistically nearly as efficient as maximum likelihood, but shows great savings in computation time.

Some Keywords: Maximum Likelihood; Pseudo-likelihood; Unequal cluster size.

1 Introduction

It is common for the number of study units in a design, the sample size, to be fixed *a priori*. However, there are many exceptions to this. Molenberghs *et al.* (2014) provides an overview of such designs. Some examples are sequential trials, for which the sample size is determined by a stopping rule,

missing data, and censored time-to-event data. The focus of this paper is hierarchical data where independent replicates take the form of compound units, generically termed *clusters*. Clustering can occur in many settings, for example, longitudinal data, toxicology (Aerts *et al.*, 2002), cluster randomized trials and, more generally, multi-level designs. The sizes of the resulting clusters can be fixed by design, or may be random. Random sample sizes can be due to missingness in the data, governed by a stochastic mechanism. And, in some cases, the cluster sizes may be associated with the outcomes, often termed 'informative cluster sizes', see for example Williamson, Datta, and Satten (2003); Benhin, Rao, and Scott (2005); Hoffman, Sen, and Weinberg (2001); Cong, Yin, and Shen (2007); Chiang and Lee (2008); Wang, Kong, and Datta (2011); Aerts *et al.* (2011). However, our interest here is not on informative cluster sizes, but rather unequal cluster sizes that are themselves independent of observed and unobserved outcomes. We concentrate solely on the non-constant nature of the cluster sizes, for which joint modelling of outcomes and cluster size is not required.

Hermans *et al.* (2016) considered the setting of normally distributed clustered data with a compound-symmetry (CS) structure for the dependency, that is, a three-parameter multivariate normal model with a common mean μ, a common variance $\sigma^2 + d$, and a common covariance d. Hermans *et al.* (2016) showed that, unless the clusters are of the same size, the sufficient statistics are incomplete and the maximum likelihood estimators (MLEs) do not have closed-form solutions. By contrast, if the clusters are all the same size, and hence also within a subset of clusters of the same size, a closed form solution does exist for the MLEs: sufficient statistics are complete and the estimators are minimum variance unbiased. Molenberghs *et al.* (2011) studied the CS case and proposed a pseudo-likelihood split-sample approach, as follows. The original sample is divided into subsamples. Maximum likelihood estimation is separately applied to each subsample and the resulting subsample-specific estimators are averaged using appropriate weights. Appropriate measures of precision for the combined estimators are then obtained. In the current setting it is natural, with this approach, to define the subsamples according to the cluster size (Hermans *et al.*, 2016). When the number of clusters and/or the cluster sizes are very large, standard iterative MLE computation times can become prohibitive. By contrast, the non-iterative nature of the split sample approach can lead to much lower computation times. For the CS setting, Hermans *et al.* (2016) show that weights proportional to the cluster sizes perform very well for combining the individual estimators.

Although the CS covariance structure is a natural model for settings that exhibit within-cluster

symmetry, other settings, such as longitudinal designs, need to be handled. For these we might consider the first-order autoregressive, AR(1), structure, where it is assumed that the correlation between two measurements changes exponentially over time, that is, $\sigma_{ij} = \sigma^2\rho^{|i-j|}$. This implies that the variance of the measurements is a constant σ^2 and the covariance decreases with increasing time lag. In this paper, we apply the split-sample method to the normal AR(1)-model, which has three parameters, a common mean μ, a common variance σ^2, and correlation parameter ρ. An important question will be the appropriate choice of weights in such a setting.

As a motivating example, we consider five clinical trials in schizophrenia. These data were collected from five double-blind randomized clinical trials to compare the effects of two treatments for chronic schizophrenia: risperidone and conventional antipsychotic agents. Subjects who received doses of risperidone (4–6 mg/day) or an active control (haloperidol, perphenazine, zuclopenthixol) have been included in the analysis.

Patients were clustered within country, and longitudinal measurements were made on each subject over time. The number of patients ranges from 9 tot 128 per country with a total of 2039. The positive and negative syndrome scale (PANSS) was used to asses the global condition of a patient. This scale is constructed from 30 items, each taking values between 1 and 7, giving an overall range of 30 to 210. PANSS provides an operationalized, drug-sensitive instrument, which is useful for both typological and dimensional assessment of schizophrenia. Depending on the trial, treatment was administered for a duration of 48 weeks with at most 12 monthly measurements. Because not all subjects received treatment at the same time points and, not the same amount, the final dataset is unbalanced.

This dataset was also analyzed from a surrogate markers point perspective in Alonso *et al.* (2004). The focus here is on the treatment effect, accommodating the longitudinal nature of the response.

The rest of paper is organised as follows. In Section 2 the model formulation is given. In Section 3 the estimators for a single constant cluster size are presented. The (in)completeness property is outlined in Section 4, and in Section 5 various weighting schemes for clusters of unequal size are explored. In Section 6, a simulation study is described for the investigation of the performance of the suggested weights and the data are analysed in Section 7. The closing discussion is presented in Section 8. Some additional background material is available in a separate, web-based appendix[1].

[1]Available at https://ibiostat.be/online-resources.

2 Model Formulation

Suppose that there is a sample of N independent clusters, among which K different cluster sizes n_k ($k = 1, \ldots, K$) can be distinguished. Let the multiplicity of cluster size n_k be c_k. The total number of clusters is then $N = \sum_{k=1}^{K} c_k$. Denote the outcome vector for the ith ($i = 1, \ldots, c_k$) replicate among the clusters of size n_k by $\boldsymbol{Y}_i^{(k)}$.

All models considered in this paper will be versions of the following general linear mixed model:

$$\boldsymbol{Y}_i^{(k)} | \boldsymbol{b_i}^{(k)} \sim N(X_i^{(k)}\boldsymbol{\beta} + Z_i^{(k)}\boldsymbol{b_i}^{(k)}, \Sigma_i^{(k)}), \quad (1)$$

$$\boldsymbol{b_i}^{(k)} \sim N(0, D), \quad (2)$$

where $\boldsymbol{\beta}$ is a vector of fixed effects, and $X_i^{(k)}$ and $Z_i^{(k)}$ are design matrices. In what follows, we consider an AR(1) covariance structure, in which case the term $Z_i^{(k)}\boldsymbol{b_i}^{(k)}$ drops from (1), while $\Sigma_i^{(k)} = \sigma^2 C_{n_k}$, with entry (r, s) equal to $\rho^{|r-s|}$. For ease of exposition, the mean structure will often be taken to be $\mu\mathbf{1}_{n_k}$, with $\mathbf{1}_{n_k}$ an n_k column vector of ones.

Note that this is very different from the a so-called balanced conditionally independent model. The contrast between this setting and the AR(1) model holds some useful insight. The interested reader can find details about this in the separate Appendix A.

3 Estimators

We begin by assuming that there is only one cluster size occurring, that is, $n_k \equiv n$ and the index k will be dropped from notation throughout this section. The resulting expressions are required for our eventual goal, clusters with variable size, which we reach in Section 5.

Again, for the present, we confine attention to clusters of constant size n. (For the purpose of identifiability we assume that there are clusters of size at least two.) Consequently, all dimension-indication subscripts n_k on matrices and vectors can be dropped until we reach Section 5. The AR(1) model of Section 2 can then be written as:

$$\boldsymbol{Y}_i \sim N\left(X_i\boldsymbol{\beta}, \Sigma = \sigma^2 C\right).$$

Because $C \equiv C(\rho)$, the parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\beta}', \rho, \sigma^2)$. When the mean is constant $\boldsymbol{\mu}_i = X_i\boldsymbol{\beta} = \mu\mathbf{1}$. It is often stated that the MLE for the AR(1) model, with a constant or more elaborate mean structure, requires numerical iteration. This is certainly the case when not all clusters are of

the same size. However, in the constant cluster size case considered here, there is a closed-form solution. Our development follows, in part, Kenward (1981).

For c clusters of length n, the kernel of the log-likelihood takes the form:

$$\ell \propto -\frac{c}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{c}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)'\Sigma^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i). \tag{3}$$

The score equation for the mean produces, as usual:

$$\widehat{\boldsymbol{\beta}} = \frac{1}{c}\sum_{i=1}^{c}(X_i'\Sigma^{-1}X_i)^{-1}(X_i'\Sigma^{-1}\boldsymbol{Y}_i). \tag{4}$$

Consider (4) for the case of a constant mean. If Σ corresponds to independence or compound-symmetry, the MLE for μ is the ordinary sample average, it does not depend on covariance parameters. For a general design $\boldsymbol{\beta}$ is estimated by the OLS estimator. However, in our AR(1) case, solving the score equations leads to:

$$\widehat{\mu} = \frac{1}{c[(n-2)(1-\rho)+2]}\sum_{i=1}^{c}\left(\sum_{j=1}^{n}Y_{ij} - \rho\sum_{j=2}^{n-1}Y_{ij}\right). \tag{5}$$

Not only does (5) depend on ρ (hence the MLE for ρ needs to be plugged in), it differs from the OLS:

$$\widetilde{\mu} = \frac{1}{cn}\sum_{i=1}^{c}\sum_{j=1}^{n}Y_{ij}. \tag{6}$$

It follows easily that, when $\rho = 0$ both estimators are the same, as it should. Interestingly, when $\rho = \pm 1$:

$$\begin{aligned}
\widehat{\mu}(\rho=+1) &= \frac{1}{c}\sum_{i=1}^{c}\frac{Y_{i1}+Y_{in}}{2}, \\
\widehat{\mu}(\rho=-1) &= \frac{1}{c(n-1)}\sum_{i=1}^{c}\left(\sum_{j=1}^{n}Y_{ij} - \frac{Y_{i1}+Y_{in}}{2}\right).
\end{aligned}$$

Turning to the score equations for the variance components, $\partial\ell/\partial\sigma^2$ leads to

$$\sigma^2 = \frac{1}{cn}\sum_{i=1}^{c}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)'C^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i). \tag{7}$$

Through C, the right-hand side depends on ρ. For ρ, we find:

$$\sigma^2 \frac{2\rho}{1-\rho^2} = \frac{1}{c(n-1)} \sum_{i=1}^{c} (\boldsymbol{y}_i - \boldsymbol{\mu}_i)' F (\boldsymbol{y}_i - \boldsymbol{\mu}_i), \tag{8}$$

with

$$F = \frac{\partial C^{-1}}{\partial \rho} = \frac{1}{(1-\rho^2)^2} \text{tridiag} \left\{ [2\rho, 4\rho, \ldots, 4\rho, 2\rho]' ; \left[-(1+\rho^2), \ldots, -(1+\rho^2) \right]' \right\}, \tag{9}$$

and with tridiag$(\boldsymbol{v}_1, \boldsymbol{v}_2)$ a tri-diagonal matrix with $\boldsymbol{v}_1$ along the main diagonal and $\boldsymbol{v}_2$ on the adjacent diagonals. Both (7) and (8) contain a summation that can be rewritten as tr$(S \cdot Q)$, with

$$S = \sum_{i=1}^{c} (\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)'$$

and Q either C^{-1} or F, as in (9), respectively. Using this formulation, and some straightforward but tedious algebra, produces:

$$f(\rho) = (n-1) S_2 \rho^3 - (n-2) R \rho^2 - (n S_2 + S_1) \rho + n R = 0, \tag{10}$$

the solution of which is the MLE $\widehat{\rho}$. Here,

$$S_1 = \sum_{j=1}^{n} s_{jj}, \qquad S_2 = \sum_{j=2}^{n-1} s_{jj}, \qquad R = \sum_{j=1}^{n-1} s_{j,j+1}. \tag{11}$$

These can be plugged into (5) to obtain $\widehat{\mu}$ and into:

$$\widehat{\sigma}^2 = \frac{1}{c} \cdot \frac{1}{1-\widehat{\rho}^2} \left(S_1 + \widehat{\rho}^2 S_2 - 2\widehat{\rho} R \right), \tag{12}$$

to obtain the MLE for σ^2.

It is easy to see that $f(\rho)$ has a single root in $[-1, 1]$. Indeed, $f(-\infty) = -\infty$, $f(+\infty) = +\infty$, $f(-1) > 0$, and $f(1) < 0$. The other two real roots are therefore in $]-\infty, -1]$ and $[1, +\infty[$. The general solution of a third-degree polynomial follows from Cardano's method. The polynomial under study was examined by Kenward (1981) who, using results of Koopmans (1942), derived an expression for the solution inside $[-1, 1]$. Alternatively, the method of Shelbey (1975) can be used.

It takes the following form. Write the polynomial symbolically as $f(\rho) = a\rho^3 + b\rho^2 + c\rho + d$, define

$$p = \frac{3ac - b^2}{3a^2}, \qquad q = \frac{2b^3 - 9abc + 27a^2d}{27a^3},$$

and further

$$C(p,q) = 2\sqrt{-\frac{p}{3}}\cos\left[\frac{1}{3}\arccos\left(\frac{3q}{2p}\sqrt{-\frac{3}{p}}\right)\right].$$

For three real roots $t_0 \leq t_1 \leq t_2$, it follows that $t_0 = C(p,q)$, $t_2 = -C(p,-q)$, and $t_1 = -t_0 - t_2$. Finally, $\widehat{\rho} = t_1 - b/(3a)$. While not as simple as the other explicit expressions for estimators, the key point is that it has a closed-form which, in turn, can be used to obtain a closed form solution for the mean and the variance, using (5) and (12), respectively. Given that it is unambiguously clear which of the three cubic solutions is the right one, no comparisons are needed, which enhances computational efficiency.

We now turn to the second derivatives in view of precision estimation. Denote by $\mathcal{I}$ the information matrix. In the usual fashion: $\mathcal{I}_{\beta\beta} = \sum_{i=1}^{c} X_i'\Sigma^{-1}X_i$. For a simple common mean μ, this becomes: $\mathcal{I}_{\mu\mu} = c\,[n-(n-2)\rho]\,/[\sigma^2(1+\rho)]$. Algebraic derivations, sketched in separate Appendix B, lead to:

$$\mathcal{I}_{\sigma^2\rho,\sigma^2\rho} = c\begin{pmatrix} \frac{n}{2(\sigma^2)^2} & -\frac{n-1}{\sigma^2}\cdot\frac{\rho}{1-\rho^2} \\ -\frac{n-1}{\sigma^2}\cdot\frac{\rho}{1-\rho^2} & \frac{(n-1)(1+\rho^2)}{(1-\rho^2)^2} \end{pmatrix}. \tag{13}$$

It is convenient to slightly change (13) to

$$\widetilde{\mathcal{I}}_{\sigma^2\rho,\sigma^2\rho} = c\begin{pmatrix} \frac{n-1}{2(\sigma^2)^2} & -\frac{n-1}{\sigma^2}\cdot\frac{\rho}{1-\rho^2} \\ -\frac{n-1}{\sigma^2}\cdot\frac{\rho}{1-\rho^2} & \frac{(n-1)(1+\rho^2)}{(1-\rho^2)^2} \end{pmatrix}, \tag{14}$$

yielding a very simple inverse:

$$\widetilde{\mathcal{I}}^{-1}_{\sigma^2\rho,\sigma^2\rho} = \frac{1}{c(n-1)}\begin{pmatrix} \frac{2(\sigma^2)^2(1+\rho^2)}{1-\rho^2} & 2\sigma^2\rho \\ 2\sigma^2\rho & 1-\rho^2 \end{pmatrix}. \tag{15}$$

4 Complete and Incomplete Sufficient Statistics

The property of central interest in this section is that of *completeness* (Casella and Berger, 2001, pp. 285–286). It means that any measurable function of a sufficient statistic, that has the zero expectation for every value of the parameter indexing the parametric model class, is the zero function almost everywhere. As has been shown in the sequential trial context, a lack of completeness does

not preclude the existence of estimators with desirable properties (Molenberghs *et al.*, 2014).

Liu and Hall (1999) established the incompleteness of the sufficient statistic for a clinical trial with a stopping rule, for the case of normally distributed endpoints. Liu *et al.* (2006) generalized this result to the exponential family. Molenberghs *et al.* (2014) and Milanzi *et al.* (2016) broadened it further to a stochastic rather than a deterministic stopping rule. They showed that, while the maximum likelihood estimator for the mean exhibits some small-sample bias and is not uniformly best, it still is consistent and asymptotically normal, in most settings, and it is a very reasonable choice.

Hermans *et al.* (2016) studied in detail the compound-symmetry model, which is essentially our model but with the AR(1) variance-covariance matrix replaced by $\sigma^2 I_{n_k} + dJ_{n_k}$, J_{n_k} being an $n_k \times n_k$ matrix of ones. They showed that, while the CS model yields a complete sufficient statistic when the cluster size is constant, this is no longer the case when at least 2 different cluster sizes occur. Rather than the definition of a complete sufficient statistic, they used a convenient characterization (Milanzi *et al.*, 2015), stating that a sufficient statistic $\boldsymbol{k}$ is complete for a parameter $\boldsymbol{\theta}$ in a exponential family model if and only if $\boldsymbol{\theta}$ cannot be transformed 1-1 to a parameterization $\boldsymbol{\eta}$ with a proper subset $\boldsymbol{\eta}_1$ such that $\boldsymbol{\eta} = [\boldsymbol{\eta}_1', \boldsymbol{\eta}_2(\boldsymbol{\eta}_1)']'$. In the type of settings that we consider, the minimal sufficient statistic is complete if it is of the same dimension as the estimator.

In what follows, we will establish completeness for the balanced conditional independence model, with the reverse holding for AR(1) model. So, in contrast to the balanced growth curve model and the compound-symmetry model with constant cluster size, an AR(1) model with constant cluster size does not allow complete sufficient statistics. This leads to some surprising results in the AR(1) case, as well as in a number of related settings of a temporal and/or spatial nature. Some of these have been alluded to in the literature of the interbellum and the early post-war period.

4.1 Balanced Conditionally Independent Model.

This model of which the estimators are spelt out in Section A, obviously admits a complete minimal sufficient statistic because the numbers of sufficient statistics (A.1)–(A.4) and estimators match (A.5)–(A.8).

4.2 AR(1) Model

The mean estimator (5) consists of two sufficient statistics:

$$K_1 = \sum_{i=1}^{c}\sum_{j=1}^{n} Y_{ij}, \qquad K_2 = \sum_{i=1}^{c}\sum_{j=2}^{n-1} Y_{ij}, \tag{16}$$

with the sufficient statistics for σ^2 and ρ spelt out in (11). In other words, the three-component vector $\boldsymbol{\theta} = (\mu, \sigma^2, \rho)'$ has a minimal sufficient statistic (K_1, K_2, S_1, S_2, R) of dimension 5, establishing incompleteness.

Even though the AR(1) model has not been studied before from the perspective of incomplete sufficient statistics, its ramifications have been mentioned in the literature. For example, as described by Martin (2006), Papadakis proposed, as early as 1937, a correction to the least-squares estimator for correlated observations arising in such settings as adjoining plots designs (Papadakis, 1937; Bartlett, 1938, 1976, 1978). The topic was also touched upon by Cochran and Bliss (1948), in the context of discriminant analysis combined with analysis of covariance. Clearly, the opportunity for such an *ad hoc* correction arises from the incompleteness. Martin (2006) and earlier authors discussing Papadakis' method refer to the somewhat unusual dependence of the mean estimator on the variance components. This parallels the property of the MLE for the mean in the AR(1) case, as in (5). Indeed, because ρ is estimated from solving a third-degree polynomial with coefficients that are functions of the sums of squares and cross-products matrix, it too is a function of such deviations. Of course, the ρ in our case is more complex than Papadakis' correction, which was more of an *ad hoc* nature, while our estimator is the solution to the likelihood equations. In essence, Papadakis' method builds a covariate from deviations observed from adjacent plots. Especially when the plots are arranged as a linear array, the connection with AR(1) is strong. Both non-iterative and iterative versions were proposed by Papadakis. In the iterative case, the covariate is re-built after every iteration, using the current value of the parameters. In more general settings, the data have a spatial layout.

In all of these cases, dependency on adjacent observations gives rise to tri-diagonal matrices, like C^{-1} in the AR(1) setting.

Cochran and Bliss (1948, p. 172) noted that the relative efficiency of the estimators with or without the use of covariance is not uniformly larger or smaller than one, but that for sufficiently large sample sizes the difference between them is small. This is entirely consistent with our findings for the AR(1) case. For Papadakis' method, the impact on bias and efficiency is described by Martin

(2006). We refer to our simulations in Section 6.

Because there is no complete minimal sufficient statistic, the MLE is not *a priori* guaranteed to be optimal. Any claims of optimality need to be demonstrated directly.

Proposition 1. *In the AR(1) model with constant mean μ and variance-covariance parameters σ^2 and ρ, and with constant cluster size, the MLE for μ is optimal (in the sense of asymptotically most efficient) and linear in the observations, with weights that depend on the parameters only through ρ.*

Note that this is not the ordinary uniform optimality. In case we demand an estimator that does not depend on the parameters at all, it cannot be uniformly more efficient than the MLE, implying that there is no such uniform estimator. The proof is given in separate Appendix B.4. This results offers the opportunity to consider estimators, based on weighting that, while not statistically fully efficient, have computational advantages such as stability (e.g., by being entirely non-iterative) and speed.

Proposition 2. *The result of Proposition 1 easily generalizes to a mean of the form $\boldsymbol{\mu} = X\boldsymbol{\beta}$, when the design is constant among clusters.*

5 Clusters Of Variable Size

Hermans *et al.* (2016) studied various weighting schemes for clusters of unequal size in the compound-symmetry case. Their work was rooted in the pseudo-likelihood and split-sample methods of Fieuws and Verbeke (2006) and Molenberghs *et al.* (2011). We will not reproduce their entire argument here, it suffices to focus on the following two-stage procedure:

1. Consider the MLE estimator for each of the K strata, defined by cluster sizes n_k and with c_k replicates. Denote these estimators generically by $\widehat{\boldsymbol{\theta}}_k$, with variance V_k.

2. Combine the $\widehat{\boldsymbol{\theta}}_k$ in an overall estimator

$$\widetilde{\boldsymbol{\theta}}^* = \sum_{k=1}^{K} A_k \widehat{\boldsymbol{\theta}}_k, \tag{17}$$

$$\text{var}(\widetilde{\boldsymbol{\theta}}^*) = \sum_{k=1}^{K} A_k V_k A_k'. \tag{18}$$

Hermans *et al.* (2016) showed that the sum of the weight matrices should be the identity matrix,

an obvious result, and considered, among others, the optimal expression:

$$A_k^{\mathsf{opt}} = \left(\sum_{m=1}^{K} V_m^{-1} \right)^{-1} V_k^{-1}. \tag{19}$$

In the AR(1) case the mean and the variance components are asymptotically independent, hence we can consider them separately. Of course, the variance components are still dependent among them.

For a general mean structure $\boldsymbol{\mu}_i^{(k)} = X_i^{(k)}\boldsymbol{\beta}$, $V_k = \sum_{i=1}^{c_k} X_i^{(k)} \Sigma_k^{-1} X_i^{(k)'}$, and the above can be applied. Note that $\Sigma_k = \sigma^2 C_k$ with C_k the AR(1) correlation matrix of dimension n_k.

Using optimal weights the $\boldsymbol{\beta}$ coefficients can then be estimated by:

$$\widetilde{\boldsymbol{\beta}} = \left(\sum_{k=1}^{K} \sum_{i=1}^{c_k} X_i^{(k)'} C_k^{-1} X_i^{(k)} \right)^{-1} \left(\sum_{k=1}^{K} \sum_{i=1}^{c_k} X_i^{(k)'} C_k^{-1} Y_i^{(k)} \right). \tag{20}$$

In the special case that the mean is constant, all $X_i^{(k)}$ are vectors of ones and then

$$\mathsf{var}(\widehat{\mu}_k) = v_k = \frac{\sigma^2(1+\rho)}{c_k} \cdot \frac{1}{[n_k - (n_k - 2)\rho]}. \tag{21}$$

The optimal weight is then

$$a_k = \frac{c_k \left[n_k - (n_k - 2)\rho\right]}{\sum_{m=1}^{K} c_m \left[n_m - (n_m - 2)\rho\right]}. \tag{22}$$

It is insightful to consider (22) in a few special cases:

$$\begin{aligned} a_k(\rho = 0) &= \frac{c_k n_k}{\sum_{m=1}^{K} c_m n_m}, \\ a_k(\rho = 1) &= \frac{c_k}{\sum_{m=1}^{K} c_m}, \\ a_k(\rho = -1) &= \frac{c_k (n_k - 1)}{\sum_{m=1}^{K} c_m (n_m - 1)}. \end{aligned}$$

Note that, even though the matrix C is singular for $\rho = \pm 1$, by taking limits, expressions can be found also for these cases. For every $\rho \neq 1$, it follows that if the n_k are sufficiently large: $a_k \approx a_k(\rho = 0)$. This implies that in a broad range of cases, except when $\rho = 1$ (or very close to it), the weights are proportional to the number of observations in a stratum, i.e., $c_k n_k$. We term these *size-proportional weights*. When $\rho = 1$ (a case where AR(1) and compound-symmetry coincide), the weights are instead *proportional*, that is, proportional to c_k.

How well the approximation works is seen in a few special cases. When $\rho = 0.5$, $a_k \propto c_k(n_k + 2)$; for $\rho = 0.9$ this becomes $a_k \propto c_k(n_k + 18)$; finally for $\rho = 0.99$, we find $a_k \propto c_k(n_k + 198)$. Thus,

for larger correlations, the size-proportionally matches clusters of sizes much larger than actually observed. But again, in practice, it is convenient and reasonable to operate under size-proportionality.

When estimating the variance of

$$\widetilde{\mu} = \sum_{k=1}^{K} a_k \mu_k, \tag{23}$$

using (22), the fact that the weights depend on ρ_k needs to be taken into account. Applying the delta method to (23), and using the variance expressions in both (21) and (15), we find:

$$\text{var}(\widetilde{\mu}) = \frac{\sum_{k=1}^{K} a_k' \sigma_k^2 (1+\rho_k)}{\left(\sum_{k=1}^{K} a_k'\right)^2} + \frac{\sum_{k=1}^{K} \left[c_k(n_k-2)\sum_{m=1}^{K} a_m'(\mu_k-\mu_m)\right]^2 \frac{1-\rho_k^2}{c_k(n_k-1)}}{\left(\sum_{k=1}^{K} a_k'\right)^4}. \tag{24}$$

We can plug in the stratum-specific $\widehat{\rho}_k$ and $\widehat{\sigma}_k^2$, or instead use the overall $\widehat{\rho}$ and $\widehat{\sigma}^2$. In the latter case, (24) becomes:

$$\text{var}(\widetilde{\mu}) = \sigma^2(1+\rho)\left\{\frac{1}{\sum_{k=1}^{K} a_k'}\right\} + (1-\rho^2)\left\{\frac{\sum_{k=1}^{K} \frac{c_k(n_k-2)^2}{(n_k-1)}\left[\sum_{m=1}^{K} a_m'(\mu_k-\mu_m)\right]^2}{\left(\sum_{k=1}^{K} a_k'\right)^4}\right\}. \tag{25}$$

Turning to the variance components, we start from (14), and use $V_k^{-1} c_k(n_k-1)P$ with

$$P = \begin{pmatrix} \frac{1}{2(\sigma^2)^2} & -\frac{1}{\sigma^2}\cdot\frac{\rho}{1-\rho^2} \\ -\frac{1}{\sigma^2}\cdot\frac{\rho}{1-\rho^2} & \frac{1+\rho^2}{(1-\rho^2)^2} \end{pmatrix}.$$

Now, clearly, the form of P does not matter because it does not depend on c_k and n_k, that is, it is free of stratum-specific quantities. This leads to:

$$A_k = \frac{c_k(n_k-1)}{\sum_{m=1}^{K} c_m(n_m-1)} P^{-1}P = \frac{c_k(n_k-1)}{\sum_{m=1}^{K} c_m(n_m-1)} I_2,$$

with I_2 the identity matrix of dimension 2. There are several implications. First, the two variance components have a diagonal weight matrix, implying that mean, variance, and correlation can be treated separately. Second, the variance and correlation have the same sets of weights. Third, they are identical to the weights for the mean when $\rho = -1$. Fourth, because these in themselves are similar to size-proportional weights, we can simplify calculations considerably, especially in large data sets, as follows:

1. Compute $\widehat{\mu}_k$, $\widehat{\sigma}_k^2$, and $\widehat{\rho}$, using the available closed-form expressions for the MLE.

2. Construct a weighted average of these using size-proportional weights.

Given that the MLE for unequal cluster sizes does not exist in closed form and hence requires iteration, this two-stage approach is nearly optimal, non-iterative, and hence fast.

Algebraic details on formulas can be found in separate Appendix B.5 and B.6.

6 Computational Considerations and Simulation Study

In the compound-symmetry covariance structure case it has been seen that the proportional weights perform very well. Due to a constant correlation d, additional observations within a cluster contribute increasingly less information relative to that already observed. By contrast, with an AR(1) covariance structure, the roles of c_k and n_k are quite different.

A first simulation study was carried out to compare the use of proportional and size-proportional weights with respect to changes in ρ. The number of clusters c_k is considered large, but the sizes n_k small. These have been chosen such that equal weights would become identical to the size-proportional weights. In this way we may see how proportional weights can work even worse in some cases. In addition, optimal weights and full likelihood were considered in the comparison. The results are presented together with those obtained for the compound-symmetry case by Hermans *et al.* (2016).

For the simulation we took: $c_1 = 500$, $c_2 = 250$, $c_3 = 250$, $c_4 = 500$, and $n_1 = 5$, $n_2 = 10$, $n_3 = 10$, $n_4 = 5$. Parameters are set as $\mu = 0$, $\sigma = 2$ and $\rho \in \{0.01, 0.2, 0.5, 0.8, 0.9, 0.95, 0.99\}$. The data are generated 100 times and the model is fitted using PROC MIXED in SAS (for a single overall intercept).

The results show that, in contrast to the CS case, with an AR(1) covariance structure the size-proportional weights give acceptable results, implying an important role for the clusters sizes, the n_k's. Proportional weights perform more poorly than equal weights.

The iterative optimal weights will converge in just one iteration (for both CS and AR(1)), which means that iterative optimal weights are nothing but approximated optimal weights. Instead of using $\widehat{\theta}_k$ one could also use $\tilde{\theta}$, obtained by using some proper weighting.

In the CS case the iterative optimal weights mainly converge to proportional weights, but with AR(1), they converge to neither proportional nor size-proportional weights. They rather converge to approximated optimal weights which are obtained by substituting the unknown parameter by its estimate using size-proportional weights.

It is observed that, for $\widehat{\mu}$ and $\widehat{\sigma}^2$, using $\tilde{\theta}$ in optimal weights does not increase the variance to a noticeable degree, but the effect for $\widehat{\rho}$ is dramatic. Though it seems that for a larger ρ this effect is diminished. Finding the proper variances when using $\tilde{\theta}$ to approximate optimal weights could be advantageous.

A second simulation study was conducted to compare computation time for closed form solutions to numerical solutions. Using closed form solutions reduces computation time significantly. Details can be found in the separate Appendix C.

7 Application: Clinical Trials in Schizophrenia

The data, introduced in Section 1, are analysed here. The active treatments are: risperidone, haloperidol, perphenazine, and zuclopernthixol. We included for analysis patients with at least one follow-up measurement. Table A.10 shows the number of patients participating in each trial for all different time patterns in receiving the treatments. As one may see, there are 26 different time patterns, therefore, the final dataset is unbalanced. This makes it suitable for examining the performance of sample splitting according to the cluster size.

For the sake of simplicity, we just take the most frequent cluster pattern for each cluster size. The model used to study the effect of the treatments on the response variable is as follows:

$$Y_{ij} = \mu + \alpha_i + \beta t_{ij} + (\alpha\beta)_{ij} + \epsilon_{ij},\ i = 1, \ldots, 4,\ j = 1, \ldots, n,\ \epsilon_{ij} \sim N_n(0, R), \qquad (26)$$

with $R_{\ell m} = \sigma^2 \rho^{|\ell - m|}$ as elements of R, β as the time effect, α_i as the treatment effect, $(\alpha\beta)_{ij}$ as the time and treatment interaction, and μ as the overall mean. For dummy coding, perphenazine has been taken as the reference treatment level.

Table 2 shows the treatment levels which appear in the different splits. Not all the treatments are present in each split. In other words, not all the splits are contributing to the estimation of every parameter. This fact should be taken into account for constructing the weights. For example, for estimating levomepromazine effect, just the first two splits are contributing, therefore, we have $(c_1 = 142, n_1 = 2)$ and $(c_2 = 143, n_2 = 3)$, which give proportional weights as $(0.498, 0.502)$, and the size-proportional weights as $(0.398, 0.602)$.

Table 3 shows the parameter estimates using sample splitting with proportional and size-proportional weights, compared to the full sample data. Note that, while the point estimates, for example for Zuclopenthixol, differ even in signs, this has to be seen against the background of the precision

Table 1: *PANSS data. Number of clusters in each trial for each cluster pattern. The pattern consists of the numbers representing the months after starting point for which a PANSS score is available.*

		Trial					
n	Pattern	FIN-1	FRA-3	INT-2	INT-3	INT-7	Total
	$(0,1)$	17	8	71	43	3	142
2	$(0,2)$	0	0	2	0	1	3
	$(0,4)$	0	0	1	0	0	1
	$(0,1,2)$	8	4	83	41	7	143
3	$(0,2,4)$	0	0	2	0	0	2
	$(0,1,4)$	1	0	3	1	0	5
	$(0,1,2,4)$	11	0	85	66	5	167
	$(0,2,4,6)$	0	0	1	0	1	2
4	$(0,2,4,8)$	0	0	1	0	0	1
	$(0,1,2,6)$	0	0	3	0	0	3
	$(0,1,2,3)$	0	4	1	0	0	5
	$(0,1,3,6)$	0	1	0	0	0	1
	$(0,2,6,8)$	0	0	0	0	1	1
	$(0,1,2,4,6)$	58	0	85	35	6	184
	$(0,1,2,4,8)$	0	0	8	0	1	9
	$(0,1,4,6,8)$	0	0	6	0	0	6
5	$(0,1,2,6,8)$	0	0	8	0	0	8
	$(0,2,4,6,8)$	0	0	3	0	2	5
	$(0,2,4,8,12)$	0	0	1	0	0	1
	$(0,1,2,3,4)$	0	44	0	0	0	44
	$(0,1,3,4,5)$	0	1	0	0	0	1
	$(0,1,2,4,6,8)$	0	0	986	240	74	1300
	$(0,1,4,6,8,10)$	0	0	1	0	0	1
6	$(0,1,2,6,8,12)$	0	0	1	0	0	1
	$(0,1,2,4,6,10)$	0	0	1	0	0	1
	$(0,1,2,4,5,6)$	0	0	2	0	0	2

estimates; their confidence intervals largely overlap.

As mentioned previously, these data are assembled from 5 trials. It might be useful to include the trial and its interaction with the variables already in the model (26) to control for the trial effect:

$$Y_{ijk} = \mu + \tau_i + \alpha_j + \beta t_{ij} + (\tau\alpha)_{ij} + (\tau\beta)_{ik} + (\alpha\beta)_{jk} + (\tau\alpha\beta)_{ijk} + \epsilon_{ijk},$$
$$i = 1,\ldots,5,\ j = 1,\ldots,4,\ k = 1,\ldots,n,\ \epsilon_{ijk} \sim N_n(0,R), \quad (27)$$

with $R_{\ell m} = \sigma^2\rho^{|\ell-m|}$ as elements of R, β as the time effect, α_j as the treatment effect, τ_i as the trial effect, $(\tau\alpha)_{ij}$ as the trial and treatment interaction, $(\tau\beta)_{jk}$ as the trial and time interaction, $(\alpha\beta)_{jk}$ as the treatment and time interaction, $(\tau\alpha\beta)_{ijk}$ as the three-way trial, treatment and time

Table 2: *PANSS data. Contributing splits in estimating each parameter. A checkmark signifies that a split contributes, a hyphen the reverse.*

Parameter	Split 1	Split 2	Split 3	Split 4	Split 5
Intercept	✓	✓	✓	✓	✓
time	✓	✓	✓	✓	✓
haloperidol	✓	✓	✓	✓	✓
levomepromazine	✓	✓	-	-	-
risperidone	✓	✓	✓	✓	✓
zuclopenthixol	✓	✓	✓	✓	-
t*haloperidol	✓	✓	✓	✓	✓
t*levomepromazine	✓	✓	-	-	-
t*risperidone	✓	✓	✓	✓	✓
t*zuclopenthixol	✓	✓	✓	✓	-
correlation ρ	✓	✓	✓	✓	✓
variance σ^2	✓	✓	✓	✓	✓

interaction, and μ as the overall mean.

Table 4 shows the estimates for the parameters of interest in this model.

Justification of the chosen model and further details as confidence limits of the tabulated estimates can be found in separate Appendix D.

8 Concluding Remarks

As an extension to the normal-compound symmetry model, discussed in Hermans *et al.* (2016), the normal AR(1) model was studied in the light of computationally effective estimation for clustered data with unequal cluster sizes.

For constant cluster size there are closed-form solutions but no complete minimal sufficient statistics. However the MLE is shown to be optimal, with weights depending on ρ for the mean. Returning to unequal cluster sizes, there are, in general, no closed form solutions. But again estimators have been obtained using a two-stage procedure. Estimators are calculated separately within each stratum (typically defined by cluster size) and combined in an overall estimator. Both theoretical and simulation results show excellent performance of the size-proportional weights, that is through weighting according to the number of measurements in a cluster ($c_k \cdot n_k$), rather than the number of clusters c_k in a subsample, that is, proportional weights. By contrast, the latter are a good choice for the compound-symmetry structure. Under AR(1) they are worse than equal weights. Approximate optimal weights can also be used, but this leads to an estimate of ρ with a large sample variance. In practice, it is convenient and appropriate to use size-proportional weights; these are

Table 3: *PANSS data. Estimating fixed effects and variance components and the standard deviations of these estimates using sample splitting (combined with proportional (Prop.) and size-proportional (Size.Prop.) weights) and full likelihood. The model used in here is without trial effect (26).*

Effect	Par.	Prop.	Size Prop.	Full
Intercept	μ	89.218 (3.036)	88.167 (2.956)	88.532 (2.965)
Haloperidol	α_1	-1.916 (3.254)	-1.868 (3.191)	-0.140 (3.181)
Levomepromazine	α_2	11.823 (14.155)	8.402 (14.366)	32.018 (9.729)
Risperidone	α_3	-1.474 (3.079)	-0.812 (3.000)	-0.481 (3.009)
Zuclopenthixol	α_4	-1.926 (7.245)	0.146 (7.216)	2.647 (4.187)
time	β	-3.047 (1.057)	-2.890 (0.613)	-2.928 (0.447)
time×haloperidol	$(\alpha\beta)_1$	2.146 (1.108)	1.568 (0.652)	1.068 (0.482)
time×levomepromazine	$(\alpha\beta)_2$	6.466 (9.006)	6.924 (8.668)	3.350 (4.501)
time×risperidone	$(\alpha\beta)_3$	1.831 (1.070)	1.243 (0.621)	0.842 (0.454)
time×zuclopenthixol	$(\alpha\beta)_4$	1.551 (3.609)	1.103 (2.655)	0.533 (0.743)
Correlation	ρ	0.805 (0.006)	0.818 (0.005)	0.825 (0.005)
Variance	σ^2	419.782 (10.202)	412.850 (10.018)	429.611 (10.363)

parameter free and simple to use. Simulations show, that in certain large settings, computation time can be 1000 times faster than with standard maximum likelihood.

There are missing observations in the PANSS data set. One might therefore consider possible dependencies between cluster size and the outcomes themselves. To handle such informative cluster sizes it might be of interest to extend the current methodology of this paper to a joint model including cluster size. This is a topic for further research.

For non-normal data, no corresponding closed-form formulations are possible. While gains will be less, there might still be computational advantages, in terms of time and stability, in analyzing the data in cluster-size dependent strata, followed by weighting the so-obtained estimates.

Acknowledgments

Financial support from the IAP research network #P7/06 of the Belgian Government (Belgian Science Policy) is gratefully acknowledged. The research leading to these results has also received funding from the European Seventh Framework programme FP7 2007–2013 under grant agreement Nr. 602552. We gratefully acknowledge support from the IWT-SBO ExaScience grant.

Table 4: *PANSS data. Estimating fixed effects and variance components and the standard deviations of these estimates using sample splitting (combined with proportional (Prop.) and size-proportional (Size.Prop.) weights) and full likelihood. The model used in here is with trial effect (27).*

Effect	Par.	Prop.	Size Prop.	Full
Intercept	μ	89.217 (3.016)	88.165 (2.949)	88.529 (2.950)
Haloperidol	α_1	2.249 (5.239)	1.491 (5.053)	5.878 (4.779)
Levomepromazine	α_2	-9.213 (22.578)	-12.044 (21.761)	6.673 (15.611)
Risperidone	α_3	2.353 (4.542)	2.956 (4.216)	3.132 (4.107)
Zuclopenthixol	α_4	-2.135 (11.617)	-0.877 (11.509)	3.144 (5.845)
time	β	-3.047 (1.049)	-2.890 (0.610)	-2.929 (0.446)
time×haloperidol	$(\alpha\beta)_1$	2.170 (1.835)	1.294 (1.056)	0.623 (0.738)
time×levomepromazine	$(\alpha\beta)_2$	16.104 (15.080)	17.287 (13.763)	13.812 (6.923)
time×risperidone	$(\alpha\beta)_3$	1.766 (1.716)	0.794 (0.947)	0.176 (0.613)
time×zuclopenthixol	$(\alpha\beta)_4$	5.218 (5.746)	2.041 (4.188)	0.326 (1.027)
Correlation	ρ	0.804 (0.006)	0.818 (0.005)	0.824 (0.005)
Variance	σ^2	416.190 (10.139)	410.819 (10.006)	425.741 (10.257)

References

Aerts, M., Geys, H., Molenberghs, G., and Ryan, L. (2002). *Topics in Modelling of Clustered Data*. London: Chapman & Hall.

Aerts, M., Faes, C., Hens, N., Loquiha, O., and Molenberghs, G. (2011). Incomplete clustered data and non-ignorable cluster size. In: Conesa, D., Forte, A., López-Quílez, A. and Muñoz, F. (Eds.), Proceedings of the 26th International Workshop on Statistical Modelling, València, Spain, 35–40.

Alonso, A., Geys, H., Molenberghs, G., and Kenward, M.G. (2004). Validation of surrogate markers in multiple randomized clinical trials with repeated measurements: canonical correlation approach. *Biometrics*, **60**, 845–853.

Bartlett, M.S. (1938). The approximate recovery of information from field experiments with large blocks. *Journal of Agricultural Science*, **28**, 418–427.

Bartlett, M.S. (1976). *The Statistical Analysis of Spatial Pattern.* London: Chapman & Hall.

Bartlett, M.S. (1978). Further analysis of spatial patterns: a re-examination of the Papadakis method of improving the accuracy of randomized block experiments. *Supplements Advances in Applied Probability*, **10**, 133–143.

Benhin, E., Rao, J.N.K., and Scott, A.J. (2005). Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes. *Biometrika*, **92**, 435–450.

Casella, G. and Berger, R.L. (2001). *Statistical Inference*. Pacific Grove: Duxbury Press.

Chiang, C-T., and Lee, K-Y. (2008). Efficient estimation methods for informative cluster size data. *Statistica Sinica*, **18**, 121–133.

Cochran, W.G. and Bliss, C.I. (1948). Discriminant function with covariance. *Annals of Mathematical Statistics*, **19**, 151–176.

Cong, X-J., Yin, G., and Shen, Y. (2007). Marginal analysis of correlated failure time data with informative cluster sizes. *Biometrics*, **63**, 663–672.

Fieuws, S. and Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modelling of multivariate longitudinal profiles. *Biometrics*, **62**, 424–431.

Hermans, L., Molenberghs, M., Aerts, M., Kenward, M.G and Verbeke, G. (2017). A tutorial on the practical use and implication of complete sufficient statistics. *Submitted for publication*.

Hermans, L., Nassiri, V., Molenberghs, G., Kenward, M.G., Van der Elst, W., Aerts, M., and Verbeke, G. (2016). Clusters with unequal size: maximum likelihood versus weighted estimation in large samples. *Submitted for publication.*

Hoffman, E.B., Sen, P.K., and Weinberg, C.R. (2001). Within-cluster resampling. *Biometrika*, **88**, 1121–1134.

Kenward, M.G. (1981). *An Investigation of Certain Methods for the Analysis of Repeated Measurements.* Reading, UK: Unpublished PhD thesis.

Koopmans, T. (1942). Serial correlation and quadratic forms in normal variables. *Annals of Mathematical Statistics*, **13**, 14–33.

Lange, N. and Laird, N.M. (1989). The effect of covariance structure on variance estimation in balanced growth-curve models with random parameters. *Journal of the American Statistical Association*, **84**, 241–247.

Liu, A. and Hall, W.J. (1999). Unbiased estimation following a group sequential test. *Biometrika*, **86**, 71–78.

Liu, A., Hall, W.J., Yu, K.F., and Wu, C. (2006). Estimation following a group sequential test for distributions in the one-parameter exponential family. *Statistica Sinica*, **16**, 165–81.

Martin, R.J. (2006). Papadakis method. In: *Encyclopedia of Statistical Science*, **9**.

Milanzi, E., Molenberghs, G., Alonso, A., Kenward, M.G., Tsiatis, A., Davidian, M., and Verbeke, G. (2015). Estimation after a group sequential trial. *Statistics in Biosciences*, **7**, 187–205.

Milanzi, E., Molenberghs, G., Alonso, A., Kenward, M.G., Verbeke, G., Tsiatis, A.A., and Davidian, M. (2016). Properties of estimators in exponential family settings with observation-based stopping rules. *Journal of Biometrics & Biostatistics*, **7**, 272.

Molenberghs, G., Kenward, M.G., Aerts, M., Verbeke, G., Tsiatis, A.A., Davidian, M., Rizopoulos, D. (2014). On random sample size, ignorability, ancillarity, completeness, separability, and degeneracy: sequential trials, random sample sizes, and missing data. *Statistical Methods in Medical Research*, **23**, 11–41.

Molenberghs, G., Verbeke, G. and Iddi S. (2011). Pseudo-likelihood methodology for partitioned large and complex samples. *Statistics and probability letters*, **81**, 892–901.

Papadakis, J.S. (1937). Méthodes statistiques pour des expériences sur champ. *Bulletin de l'Institut pour Amélioration des Plantes à Salonique*, **23**.

Shelbey, S. (1975). *CRC Standard Mathematical Tables*. Boca Raton: CRC Press.

Verbeke, G. and Fieuws, S. (2007) The effect of miss-specified baseline characteristics on inference for longitudinal trends in linear mixed models. *Biostatistics*, **8**, 772–783.

Wang, M., Kong, M., and Datta, S. (2011). Inference for marginal linear models for correlated longitudinal data with potentially informative cluster sizes. *Statistical Methods in Medical Research*, **20**, 347–367.

Williamson, J.M., Datta, S., and Satten, G.A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics*, **59**, 36–42.

Fast, Closed-form, and Efficient Estimators for Hierarchical Models with AR(1) Covariance and Unequal Cluster Sizes

Lisa Hermans[1] **Vahid Nassiri**[2] **Geert Molenberghs**[1,2]

Michael G. Kenward[3] **Wim Van der Elst**[4,1] **Marc Aerts**[1]

Geert Verbeke[2,1]

[1] *I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium*

[2] *I-BioStat, KU Leuven, B-3000 Leuven, Belgium*

[3] *Luton, United Kingdom*

[4] *Janssen Pharmaceutica, B-2340 Beerse, Belgium*

Appendix

A The Balanced Conditionally Independent Model

In this case, one imposes the following structure on (1):

- $X_i^{(k)}$ can be rewritten in terms of a first matrix that imposes structure between clusters (e.g., treatment effect), termed $A_i^{(k)}$, and a second one that imposes structure within clusters (e.g., time evolution), $T_i^{(k)\prime} = (Z_i^{(k)\prime}, Q_i^{(k)\prime})'$.
- The matrices $A_i^{(k)}$, $Z_i^{(k)}$, and $Q_i^{(k)}$ are constant among all clusters of size n_k.
- The matrix $\Sigma_i^{(k)} = \sigma^2 I_{n_k}$.

This is the general, balanced growth-curve model as studied by Lange and Laird (1989) and Verbeke and Fieuws (2007). Building on their development, we will now derive sufficient statistics and associated maximum likelihood estimators for the parameters in this model. This can be expressed

$$\boldsymbol{Y} = A(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \begin{pmatrix} Z \\ Q \end{pmatrix} + BZ + \boldsymbol{\varepsilon}.$$

Here, $\boldsymbol{Y}$ is an $N \times n$ matrix stacking the outcomes of all clusters of size c, A, Z, and Q group the designs mentioned in Section 2, the vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ contain the fixed effects, B contains N rows

of length q, representing the q-dimensional random-effects vector, and $\boldsymbol{\varepsilon}$ shares its dimensions with $\boldsymbol{Y}$.

Now, define K the projection matrix such that $K'K = I_{r-q}$, for an appropriate dimension r, and $ZK = 0$. Then, set $P = QK$ and consider the projection model:

$$\boldsymbol{Y}_1 \equiv YK = A\boldsymbol{\beta}_2 P + \boldsymbol{\varepsilon}K.$$

The variance of a cluster is $\sigma^2 I_{r-q}$. Next, define H such that $H'H = I_q$ and $QH = 0$. A second projection model emerges:

$$\boldsymbol{Y}_2 \equiv YH = A\boldsymbol{\beta}_1 + B + \boldsymbol{\varepsilon}H.$$

The variance of a cluster is $\sigma^2 I_q + H'DH$, with D the variance-covariance matrix of the vector of random effects. Importantly, projections $\boldsymbol{Y}_1 \perp \boldsymbol{Y}_2$.

Conventional algebra leads from these to the following set of sufficient statistics:

$$T_1 = (A'A)^{-1}A'\boldsymbol{Y}_1 P'(PP')^{-1}, \tag{A.1}$$

$$T_2 = \text{tr}\left\{\boldsymbol{Y}_1'\left[I - A(A'A)^{-1}A'\right]\boldsymbol{Y}_1\right\}, \tag{A.2}$$

$$T_3 = (A'A)^{-1}A'\boldsymbol{Y}_2, \tag{A.3}$$

$$T_4 = \boldsymbol{Y}_2'\left[I - A(A'A)^{-1}A'\right]\boldsymbol{Y}_2. \tag{A.4}$$

Sufficient statistics (A.1)–(A.4) lead to the maximum likelihood estimators:

$$\widehat{\boldsymbol{\beta}}_1 = T_1, \tag{A.5}$$

$$\widehat{\boldsymbol{\beta}}_2 = T_3, \tag{A.6}$$

$$\widehat{\sigma}^2 = \frac{1}{N(n-q)}T_2, \tag{A.7}$$

$$\widehat{D} = \frac{1}{N}T_4 - \widehat{\sigma}^2 I_q. \tag{A.8}$$

Note that the estimators for the fixed effects do not involve the variance components.

B Algebraic Derivations in the AR(1) Case

Here, we present more detailed derivations of the key algebraic expressions presented in Sections 2 and 3.

B.1 Some Useful Expressions

Consider,

$$C = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots \\ & 1 & \rho & \rho^2 & \dots \\ & & \ddots & \ddots & \ddots \\ & & & 1 & \rho \\ & & & & 1 \end{pmatrix}, \tag{A.9}$$

then,

$$\Sigma = \sigma^2 C. \tag{A.10}$$

It can be shown that:

$$\det(C) = \left(1 - \rho^2\right)^{n-1} \tag{A.11}$$

The inverse of C can be calculated as follows:

$$C^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho & \dots & 0 \\ -\rho & 1+\rho^2 & \ddots & \dots \\ & \ddots & \ddots & \ddots \\ & & 1+\rho^2 & -\rho \\ & & -\rho & 1 \end{pmatrix}, \tag{A.12}$$

as one may see C^{-1} is a symmetric-tridiagonal matrix with constant diagonal except for the outer entries, and constant first off-diagonal.

Consider:

$$C^{-1} = \frac{1}{1-\rho^2} G. \tag{A.13}$$

Then, by taking the derivative with respect to ρ:

$$\frac{\partial C^{-1}}{\partial \rho} = \frac{2\rho}{(1-\rho^2)^2} G + \frac{1}{1-\rho^2} H, \tag{A.14}$$

where, $H = \frac{\partial G}{\partial \rho}$ and has the form:

$$H = \begin{pmatrix} 0 & -1 & \dots & 0 \\ -1 & 2\rho & \ddots & \dots \\ \ddots & \ddots & \ddots & \ddots \\ & \ddots & 2\rho & -1 \\ 0 & & -1 & 0 \end{pmatrix}. \tag{A.15}$$

Also, considering the fact $CC^{-1} = I$, one can derive:

$$\frac{\partial C^{-1}}{\partial \rho} = -C^{-1}\frac{\partial C}{\partial \rho}C^{-1}. \tag{A.16}$$

B.2 The Likelihood Estimators in a Given Cluster

The likelihood function for an n_k-dimensional multivariate normal sample of size c_k has the following form:

$$L = \prod_{i=1}^{C_k} \frac{1}{|\Sigma|^{1/2}(2\pi)^{n_k/2}} \exp\left\{-\frac{1}{2}(y_i - \mu_i)'\Sigma^{-1}(y_i - \mu_i)\right\}. \tag{A.17}$$

Therefore, the non-constant terms of the log-likelihood are as follows:

$$\ell \propto -\frac{C_k}{2}\ln|\Sigma|.\frac{1}{2}\sum_{i=1}^{C_k}(y_i - \mu_i)'\Sigma^{-1}(y_i - \mu_i), \tag{A.18}$$

which considering (A.11) for AR(1):

$$|\Sigma| = \left(\sigma^2\right)^{n_k}\left(1-\rho^2\right)^{n_k-1}. \tag{A.19}$$

As a general case, if we consider the mean as linear model with the form $\mu_i = X_i\beta$, one can derive:

$$\frac{\partial \ell}{\partial \mu_i} = \Sigma^{-1}\sum_{i=1}^{C_k}(y_i - \mu_i) = 0 \Rightarrow \widehat{\beta} = (X'X)^{-1}X'y. \tag{A.20}$$

Now expanding the log-likelihood for σ^2 and ρ, we have:

$$\ell \propto -\frac{C_k}{2}n_k\ln\sigma^2 - \frac{C_k}{2}(n_k-1)\ln(1-\rho^2) - \frac{1}{2}\sum_{i=1}^{C_k}(y_i - \mu_i)'\Sigma^{-1}(y_i - \mu_i). \tag{A.21}$$

Considering $\Sigma = \sigma^2 C$ and (A.12), the derivative with respect to σ^2 is as follows:

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{C_k n_k}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{1}{(\sigma^2)^2}\sum_{i=1}^{C_k}(y_i - \mu_i)' C^{-1}(y_i - \mu_i). \tag{A.22}$$

Solving $\frac{\partial \ell}{\partial \sigma^2} = 0$ gives:

$$\widehat{\sigma}^2 = \frac{1}{C_k n_k}\sum_{i=1}^{C_k}(y_i - \mu_i)' C^{-1}(y_i - \mu_i). \tag{A.23}$$

One may notice that C^{-1} contains the parameter ρ.

Taking the derivative of (A.21) with respect to ρ gives:

$$\frac{\partial \ell}{\partial \rho} = \frac{C_k(n_k - 1)}{2}\frac{2\rho}{1-\rho^2} - \frac{1}{\sigma^2}\sum_{i=1}^{C_k}(y_i - \mu_i)'\frac{\partial C^{-1}}{\partial \rho}(y_i - \mu_i). \tag{A.24}$$

Setting $\frac{\partial \ell}{\partial \rho} = 0$ gives:

$$\widehat{\sigma}^2\frac{2\widehat{\rho}}{1-\widehat{\rho}^2} = \frac{1}{C_k(n_k-1)}\sum_{i=1}^{C_k}(y_i - \mu_i)'\frac{\partial C^{-1}}{\partial \rho}(y_i - \mu_i). \tag{A.25}$$

Solving (A.23) and (A.25) gives $\widehat{\sigma}^2$ and $\widehat{\rho}$. For any $(n_k \times n_k)$ matrix Q, $\sum_i (y_i - \mu_i)' Q (y_i - \mu_i)$ equals $\mathrm{tr}\{SQ\}$, where tr denotes the trace of a matrix, and $S = \sum_i (y_i - \mu_i)(y_i - m_i)'$. Hence, from (A.23), (A.25), (A.13), (A.14), and (A.16), one can write:

$$\begin{cases} (1-\widehat{\rho}^2)\widehat{\sigma}^2 = \frac{1}{C_k n_k}\mathrm{tr}\{SG\}, \\ (1-\widehat{\rho}^2)\widehat{\sigma}^2 = \frac{1}{C_k n_k}\mathrm{tr}\{SG\} + \frac{1-\widehat{\rho}^2}{2\widehat{\rho}}\frac{1}{C_k(n_k-1)}\frac{1}{C_k(n_k-1}\mathrm{tr}\{SH\}. \end{cases} \tag{A.26}$$

Set $g = \mathrm{tr}\{SG\}$ and $h = \mathrm{tr}\{SH\}$, it follows that

$$\frac{g}{n_k} + \frac{1-\widehat{\rho}^2}{2\widehat{\rho}}h = 0. \tag{A.27}$$

Given that both g and h are functions of ρ only, ρ can be estimated using $(A.27)$. Givenρ, one can use one of equations in (A.26) to estimate σ^2.

Let us consider some special cases. For $n_k = 2$:

$$G = \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}, \qquad H = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}.$$

Therefore, g and h can be computed as:

$$g = \mathrm{tr}\left[\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\right] = S_{11} - 2\rho S_{12} + S_{22}.$$

$$h = \mathrm{tr}\left[\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}\right] = -2S_{12}.$$

Now using (A.27):

$$\widehat{\rho}(S_{11} - 2\widehat{\rho}S_{12} + S22 + (1-\widehat{\rho}^2)(-2S_{12}),$$

which gives:

$$\widehat{\rho} = \frac{2S_{12}}{S_{11} + S_{22}}. \tag{A.28}$$

Then, using first equation in (A.26):

$$(1-\widehat{\rho}^2)\widehat{\sigma}^2 = \frac{1}{2C_k}(S_{11} - 2\widehat{\rho}S_{12} + S_{22}),$$

which gives:

$$\widehat{\sigma}^2 = \frac{S_{11} + S22}{2C_k}. \tag{A.29}$$

For $n_k = 3$:

$$g = \mathrm{tr}\left[\begin{pmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{32} & S_{33} \end{pmatrix}\begin{pmatrix} 1 & -\rho & 0 \\ -\rho & 1+\rho^2-\rho & \\ 0 & -\rho & 1 \end{pmatrix}\right] = S_{11} + S_{22} + S_{33} - 2\rho(S_{12} + S_{23}) + \rho^2 S_{22},$$

$$h = \mathrm{tr}\left[\begin{pmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{32} & S_{33} \end{pmatrix}\begin{pmatrix} 0 & -1 & 0 \\ -1 & 2\rho & -1 \\ 0 & -1 & 0 \end{pmatrix}\right] = -2(S_{12} + S_{23}) + 2\rho S_{22}.$$

Let,

$$\begin{cases} S = S_{11} + S22 + S33 \\ R = S_{12} + S_{23} \end{cases} \Rightarrow \begin{cases} g = S + \rho^2 S_{22} - 2\rho R \\ h = -2R + 2\rho S_{22} \end{cases}$$

Using (A.27):

$$2S_{22}\rho^3 - R\rho^2 - (S + 3S_{22})\rho + 3R = 0. \tag{A.30}$$

Considering the results for $n_k = 2$ and $n_k = 3$, one can calculate (A.27) for the general case $n_k = n$ as follows.

$$(n-1)\tilde{S}\rho^3 - (n-2)R\rho^2 - (n\tilde{S} + S)\rho + nR = 0 \tag{A.31}$$

with:

$$\begin{cases} S = S_{11} + \ldots + S_{nn}, \\ \tilde{S} = S_{22} + \ldots + S_{n-1,n-1}, \\ R = S_{12} + S_{23} + \ldots + S_{n-1,n}. \end{cases}$$

Then using (A.26):

$$\widehat{\sigma}^2 = \frac{1}{C_n}\frac{1}{(1-\widehat{\rho}^2)}(S + \widehat{\rho}^2\tilde{S} - 2\widehat{\rho}R). \tag{A.32}$$

For $n_k > 2$, (A.31) is a third-degree polynomial. One can show that this equation has only one root in $[-1, 1]$.

Proof. Consider:

$$\begin{aligned} f(\rho) &= (n-1)\tilde{S}\rho^3 - (n-2)R\rho^2 - (n\tilde{S} + S)\rho + nR \\ f'(\rho) &= 3(n-1)\tilde{S}\rho^2 - 2(n-2)R\rho - (n\tilde{S} + S) \\ f''(\rho) &= 6(n-1)\tilde{S}\rho - 2(n-2)R \end{aligned}$$

The discriminant of $f'(\rho)$ is as follows:

$$\Delta_{f'(\rho)} = (n-2)^2R^2 + 3(n\tilde{S} + S)(n-1)\tilde{S} \geq 0.$$

Therefore $f'(\rho)$ has no root and hence $f(\rho)$ is monotone. One may see $f'(0) \leq 0$, therefore, $f(\rho)$ is a monotonically decreasing function (I). One can show $f(1) \leq 0$ and $f(-1) \geq 0$ (II). Considering (I) and (II) together, one may conclude $f(\rho)$ must necessarily cross the horizontal line only once between $[-1, 1]$. □

This shows the unique $\widehat{\rho}$ can be easily estimated solving (A.31) using Cardano's formula (Cardano, 1979).

B.3 Hessians, Covariance Matrices, and Optimal Weights

Given the MLEs for the AR(1) covariance structure, the Hessians and covariance matrices of the MLEs can be derived. Following the general results obtained about optimal weights, they can be used to compute the exact optimal weights in the case of the AR(1) structure. As mean and variance parameters are orthogonal in the normal case, we can consider the second derivative for fixed effects and variance components separately.

B.3.1 Second derivative with respect to fixed effects

As

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{C_k} X_i' \, \Sigma^{-1}(y_i - \mu_i),$$

we have:

$$\begin{aligned} \mathrm{E}\left[\frac{\partial \ell}{\partial \beta}\left(\frac{\partial \ell}{\partial \beta}\right)'\right] &= \sum_{i=1}^{C_k} X_i' \Sigma^{-1} \mathrm{E}(y_i - \mu_i)(y_i - \mu_i)' \Sigma^{-1} X_i \\ &= \sum_{i=1}^{C_k} X_i' \Sigma^{-1} X_i. \end{aligned}$$

For the special case of just an intercept $X_i = \mathbf{1}$:

$$\mathrm{E}\left[\frac{\partial \ell}{\partial \beta}\left(\frac{\partial \ell}{\partial \beta}\right)'\right] = \sum_{i=1}^{C_k} \mathbf{1}' \Sigma^{-1} \mathbf{1} = \frac{C_k}{\sigma^2(1-\rho^2)}\left[(n_k - 2)\rho^2 - 2(n_k - 1)\rho + n_k\right]. \tag{A.33}$$

Therefore, the variance for $\widehat{\mu}$ can be computed as inverse of (A.33).

B.3.2 Second derivative with respect to variance components

To calculate the derivatives with respect to variance components rather than $\frac{\partial C^{-1}}{\partial \rho}$, we need $K = \frac{\partial C^{-1}}{\partial \rho^2}$. Using these derivatives:

$$\begin{cases} \frac{\partial}{\partial \rho} 2\left(\frac{\rho}{1-\rho^2}\right) = 2\frac{1+\rho^2}{(1-\rho^2)^2}, \\ \frac{\partial}{\partial \rho}\frac{\rho}{(1-\rho^2)^2} = \frac{1+3\rho^2}{(1-\rho^2)^3}, \\ \frac{\partial}{\partial \rho}\frac{1+\rho^2}{(1-\rho^2)^2} = \frac{2\rho(3+\rho^2)}{(1-\rho^2)^3}. \end{cases} \tag{A.34}$$

it follows that

$$\frac{\partial C^{-1}}{\partial \rho^2} = K = \frac{1}{(1-\rho^2)^3}\begin{pmatrix} 2(1+3\rho^2) & -2\rho(3+\rho^2) & & 0 \\ -2\rho(3+\rho^2) & 4(1+3\rho^2) & \ddots & \ddots \\ & \ddots & \ddots & \ddots \\ 0 & & -2\rho(3+\rho^2) & 2(1+3\rho^2). \end{pmatrix} \tag{A.35}$$

The second-derivatives are:

$$\begin{cases} \frac{\partial^2 \ell}{\partial(\sigma^2)^2} = \frac{C_k n_k}{2}\frac{1}{(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}\sum_{i=1}^{C_k}(y_i-\mu_i)'C^{-1}(y_i-\mu_i), \\ \frac{\partial^2 \ell}{\partial \rho^2} = \frac{C_k(n_k-1)(1+\rho^2)}{(1-\rho^2)^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{C_k}(y_i-\mu_i)'K(y_i-\mu_i), \\ \frac{\partial^2 \ell}{\partial \rho \partial \sigma^2} = \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{C_k}(y_i-\mu_i)'\frac{\partial C^{-1}}{\partial \rho}(y_i-\mu_i). \end{cases} \tag{A.36}$$

To construct the expected Hessian and covariance matrix, one needs to find the expectations of the expressions in (A.36).

$$\mathrm{E}\left(\frac{\partial^2 \ell}{\partial(\sigma^2)^2}\right) = -\frac{C_k n_k}{2}\frac{1}{(\sigma^2)^2}. \tag{A.37}$$

This follows from the fact that:

$$\mathrm{E}\left(\sum_{i=1}^{C_k}(y_i-\mu_i)'C^{-1}(y_i-\mu_i)\right) = C_k \mathrm{tr}\left\{\mathrm{E}\left[(y_i-\mu_i)'(y_i-\mu_i)\right]C^{-1}\right\},$$

and $\mathrm{E}\left[(y_i-\mu_i)(y_i-\mu_i)'\right] = \sigma^2 C$.

For the second derivative with respect to ρ:

$$\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \rho^2}\right] = \frac{C_k(n_k-1)(1+\rho^2)}{(1-\rho^2)^2} - \frac{C_k}{2}\mathrm{tr}\{KS\}. \tag{A.38}$$

Likewise:

$$\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \rho \partial \sigma^2}\right] = \frac{C_k}{2\sigma^2}\mathrm{tr}\left\{C\frac{\partial C^{-1}}{\partial \rho}\right\}. \tag{A.39}$$

Substituting for $\mathrm{tr}\{KS\}$ and $\mathrm{tr}\left\{C\frac{\partial C^{-1}}{\partial \rho}\right\}$ we get:

$$\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \rho \partial \sigma^2}\right] = \frac{C_k(n_k-1)}{\sigma^2}\frac{\rho}{1-\rho^2}. \tag{A.40}$$

$$\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \rho^2}\right] = -C_k(n_k-1)\frac{1+\rho^2}{(1-\rho^2)^2}. \tag{A.41}$$

Using (A.37), (A.40), and (A.41) one obtains the 2×2 Hessian matrix as follows:

$$H = -C_k \begin{pmatrix} \frac{n_k}{2(\sigma^2)^2} & -\frac{n_k-1}{\sigma^2}\frac{\rho}{1-\rho^2} \\ -\frac{n_k-1}{\sigma^2}\frac{\rho}{1-\rho^2} & (n_k-1)\frac{1+\rho^2}{(1-\rho^2)^2} \end{pmatrix}. \tag{A.42}$$

The determinant of the Hessian in (A.42) is as follows:

$$\det(H) = \frac{C_k^2(n_k-1)(n_k-(n_k-2)\rho^2)}{2(\sigma^2)^2(1-\rho^2)^2}. \tag{A.43}$$

So,

$$-H^{-1} = \frac{1}{C_k(n_k-(n_k-2)\rho^2)} \begin{pmatrix} 2(\sigma^2)^2(1+\rho^2) & 2\rho\sigma^2(1-\rho^2) \\ 2\rho\sigma^2(1-\rho^2) & \frac{n_k}{n_k-1}(1-\rho^2)^2 \end{pmatrix}. \tag{A.44}$$

The Hessian for the unbiased estimator differs slightly from its MLE counterpart:

$$\tilde{H} = -C_k \begin{pmatrix} \frac{n_k-1}{2(\sigma^2)^2} & -\frac{n_k-1}{\sigma^2}\frac{\rho}{1-\rho^2} \\ -\frac{n_k-1}{\sigma^2}\frac{\rho}{1-\rho^2} & (n_k-1)\frac{1+\rho^2}{(1-\rho^2)^2} \end{pmatrix}, \tag{A.45}$$

$$\det(\tilde{H}) = \frac{C_k^2(n_k-1)^2}{2(\sigma^2)^2(1-\rho^2)}. \tag{A.46}$$

Therefore,

$$-\tilde{H}^{-1} = \frac{1}{C_k(n_k-(n_k-2)\rho^2)} \begin{pmatrix} 2(\sigma^2)^2(1+\rho^2) & 2\rho\sigma^2(1-\rho^2) \\ 2\rho\sigma^2(1-\rho^2) & (1-\rho^2)^2 \end{pmatrix}. \tag{A.47}$$

Having the covariance matrix, one may easily find the optimal weights using

$$W_{opt.} = \frac{V_k^{-1}}{\sum_{i=1}^K V_i^{-1}} \tag{A.48}$$

The variance of an estimator obtained using the optimal weights in (A.48) can be calculated as $\left(\sum_{i=1}^K V_i^{-1}\right)^{-1}$.

B.4 Proof of Proposition 1

Proof. Consider an estimator of the form:

$$\widetilde{\mu}_\alpha = \frac{1}{c}\sum_{i=1}^{c}\sum_{j=1}^{n} \alpha_j Y_{ij}, \tag{A.49}$$

for a vector of weights $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)'$. Because the clusters are i.i.d. it is evident that the

components of $\boldsymbol{\alpha}$ do not depend on the cluster index i. Clearly, the requirement that $E(\widetilde{\mu}_\alpha) = \mu$ implies the condition

$$\sum_{j=1}^{n} \alpha_j = 1. \tag{A.50}$$

An expression of the variance of $\widetilde{\mu}_\alpha$ combined with this requirement produces the objective function:

$$Q = \sigma^2 \left(\sum_{j=1}^{n} \alpha_j^2 + 2 \sum_{j<k} \alpha_j \alpha_k \rho^{|j-k|} \right) - \lambda \left(\sum_{j=1}^{n} \alpha_j - 1 \right), \tag{A.51}$$

with λ a Lagrange multiplier. Taking the derivative of (A.51) w.r.t. $\boldsymbol{\alpha}$ leads to, after rearrangement:

$$\boldsymbol{\alpha} = \frac{\lambda}{2\sigma^2} C^{-1} \mathbf{1}.$$

Given that we have an explicit form for C^{-1}, it follows that

$$\boldsymbol{\alpha} = \frac{\lambda}{2\sigma^2(1+\rho)} \boldsymbol{\rho}^{(1)}, \tag{A.52}$$

with $\boldsymbol{\rho}^{(1)} = (1, 1-\rho, \ldots, 1-\rho, 1)'$. Combining (A.52) with constraint (A.50) leads to $\lambda = 2\sigma^2(1+\rho)/[2+(n-2)(1-\rho)]$, hence

$$\alpha = \frac{1}{[2+(n-2)(1-\rho)]} \boldsymbol{\rho}^{(1)},$$

establishing the MLE. This completes the proof.

B.5 Optimal weights in case of a general mean structure $X_i^{(k)}\boldsymbol{\beta}$

Cluster size specific expressions are:

$$\widehat{\boldsymbol{\beta}_k} = \left(\sum_{i=1}^{c_k} X_i^{(k)'} \Sigma_k^{-1} X_i^{(k)} \right)^{-1} \left(\sum_{i=1}^{c_k} X_i^{(k)'} \Sigma_k^{-1} Y_i^{(k)} \right) \tag{A.53}$$

and

$$\text{var}(\widehat{\boldsymbol{\beta}_k}) = V_k = \left(\sum_{i=1}^{c_k} X_i^{(k)'} \Sigma_k^{-1} X_i^{(k)} \right)^{-1}. \tag{A.54}$$

The combination ruleis

$$\tilde{\boldsymbol{\beta}}_k = \sum_{i=1}^{K} A_k \widehat{\boldsymbol{\beta}_k}, \tag{A.55}$$

with

$$V_k^{-1} = \frac{1}{\sigma^2} \sum_{i=1}^{c_k} X_i^{(k)'} C_k^{-1} X_i^{(k)} \tag{A.56}$$

and C_k is as described in Supplementary Materials B.1.

The first factor in (A.53) can be split into three parts:

$$(1-\rho^2) X_i^{(k)'} C_k^{-1} X_i^{(k)} = X_i^{(k)'} (1+\rho^2) I_k X_i^{(k)} \tag{A.57}$$

$$- \rho^2 X_i^{(k)'} \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & 0 & & & \vdots \\ \vdots & & \ddots & & \\ \vdots & & & 0 & 0 \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix} X_i^{(k)} \tag{A.58}$$

$$- \rho X_i^{(k)'} \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & 0 & 1 \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix} X_i^{(k)}. \tag{A.59}$$

(A.57) simplifies to $(1+\rho^2) X_i^{(k)'} X_i^{(k)}$, while (A.58) equals

$$\rho^2 \begin{pmatrix} x_{ki11}^2 + x_{kin_k1}^2 & x_{ki11} \cdot x_{ki22} + x_{kin_k1} \cdot x_{kin_k2} & \dots & x_{ki11} \cdot x_{ki1p} + x_{kin_k1} \cdot x_{kin_kp} \\ & x_{ki12}^2 + x_{kin_k2}^2 & & \\ & & \ddots & \\ & & & x_{ki1p}^2 + x_{kin_kp}^2. \end{pmatrix}. \tag{A.60}$$

Defining $\boldsymbol{X}_{i1}^{(k)} = (x_{ki11} \ \dots \ x_{ki1p})^t$ and $\boldsymbol{X}_{in_k}^{(k)} = (x_{kin_k1} \ \dots \ x_{kin_kp})^t$, (A.58) equals

$$\rho^2 [\boldsymbol{X}_{i1}^{(k)} \ 0 \dots 0 \ \boldsymbol{X}_{in_k}^{(k)}] X_i^{(k)}. \tag{A.61}$$

For the third term, define $\boldsymbol{X}_{ij}^{(k)-} = (x_{ki2j} \ \dots \ x_{kin_kj})^t$ and $\boldsymbol{X}_{ij}^{(k)+} = (x_{ki1j} \ \dots \ x_{kin_{k-1}j})^t$. As a

consequence, (A.59) will equal

$$-\rho[X_i^{(k)-'} \cdot X_i^{(k)+} + X_i^{(k)+'} \cdot X_i^{(k)-}]. \tag{A.62}$$

In summary:

$$\begin{aligned}(1-\rho^2)X_i^{(k)'}C_k^{-1}X_i^{(k)} &= (1+\rho^2)X_i^{(k)'}X_i^{(k)} \\ &\quad - \rho^2[\boldsymbol{X}_{i1}^{(k)}\ 0 \ldots 0\ \boldsymbol{X}_{in_k}^{(k)}]X_i^{(k)} \\ &\quad - \rho[X_i^{(k)-'} \cdot X_i^{(k)+} + X_i^{(k)+'} \cdot X_i^{(k)-}] \\ &\overset{\text{notation}}{=} F_{1k}.\end{aligned} \tag{A.63}$$

The second factor in (A.53), using the same notations for $Y_i^{(k)}$ as described above, can be rewritten as:

$$\begin{aligned}(1-\rho^2)X_i^{(k)'}C_k^{-1}Y_i^{(k)} &= (1+\rho^2)X_i^{(k)'}Y_i^{(k)} \\ &\quad -\rho^2 \begin{pmatrix} x_{ki11} \cdot y_{ki1} + x_{kin_k1} \cdot y_{kin_k} \\ x_{ki12} \cdot y_{ki1} + x_{kin_k2} \cdot y_{kin_k} \\ \vdots \\ x_{ki1p} \cdot y_{ki1} + x_{kin_kp} \cdot y_{kin_k} \end{pmatrix} \\ &\quad -\rho[X_i^{(k)-'} \cdot Y_i^{(k)+} + X_i^{(k)+'} \cdot Y_i^{(k)-}] \\ &\overset{\text{not.}}{=} F_{2k}.\end{aligned} \tag{A.64}$$

Combining (A.63) and (A.64) the overall estimate equals:

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_k &= \sum_{i=1}^{K} A_k \widehat{\boldsymbol{\beta}_k} \\ &= \sum_{i=1}^{K} \left(\sum_{m=1}^{K} F_{1m} \right)^{-1} F_{2k}\end{aligned} \tag{A.65}$$

B.6 Delta Method for the Mean Estimator

By plugging in ρ_k and defining $a'_k = c_k[n_k - (n_k - 2)\rho_k]$, equation (A.66) simplifies to

$$a_k = \frac{a'_k}{\sum_{m=1}^{K} a'_m}, \tag{A.66}$$

and (21) becomes

$$\text{var}(\widehat{\mu}_k) = v_k = \frac{\sigma_k^2(1+\rho_k)}{a'_k}. \tag{A.67}$$

The first derivatives equal

$$\begin{aligned}
\frac{\partial \tilde{\mu}}{\partial \mu_k} &= a_k = \frac{a'_k}{\sum_{m=1}^K a'_m}, \\
\frac{\partial \tilde{\mu}}{\partial \sigma_k^2} &= 0, \\
\frac{\partial \tilde{\mu}}{\partial \rho_k} &= \frac{-c_k(n_k-2)\sum_{m=1}^K a'_m(\mu_k-\mu_m)}{\left(\sum_{m=1}^K a'_m\right)^2},
\end{aligned} \tag{A.68}$$

and these can be combined using the delta method, resulting in (24):

$$\begin{aligned}
\text{var}(\tilde{\mu}) \;=\; & \sum_{i=1}^K \frac{a_k'^2}{\left(\sum_{k=1}^K a'_k\right)^2} \cdot \frac{\sigma_k^2(1-\rho_k^2)}{a'_k} \\
& + \frac{\sum_{k=1}^K \left[c_k(n_k-2)\sum_{m=1}^K a'_m(\mu_k-\mu_m)\right]^2}{\left(\sum_{k=1}^K a'_k\right)^4} \cdot \frac{1-\rho_k^2}{c_k(n_k-1)}.
\end{aligned}$$

B.7 Calculating $\widehat{\rho}$ and $\widehat{\sigma}^2$ in R

In this section, we consider the implementation of the calculations for the variance components in R. This can be done with a few simple lines of code. For fixed $C_k = C$ and $n_k = n$, and given the data y, the function `est.ar1` estimates the variance components and provides a plot for the third-degree polynomial in (A.31). This visually underscores that there is only one root in $[-1, 1]$. Figure A.1 shows (A.31) for 10 simulated data sets; clearly, there is a single root only in $[-1, 1]$. For convenience, the R code is given in Supplementary Materials E. Other functions to find variances and iterated optimal weights are also available.

C Details on Additional Simulations

C.1 Simulations with Proportional and Size-proportional Weights

Here we consider $C_1 = 500, C_2 = 250, C_3 = 250, C_4 = 500$, and $n_1 = 5, n_2 = 10, n_3 = 10, n_4 = 5$. Parameters are set to $\mu = 0$, $\sigma = 2$ and $\rho = (0.1, 0.5, 0.8)$. The data are generated 100 times and the model is fitted using PROC MIXED in SAS (for a single overall intercept). For combining the

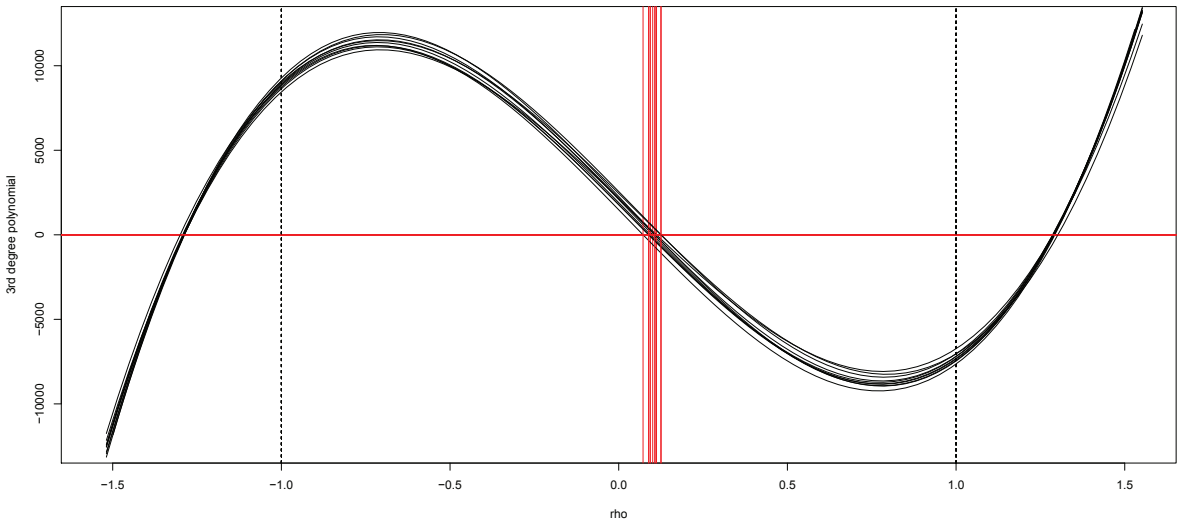

Figure A.1: *The third degree polynomial in (A.31) for* 10 *different generated data. The red veertical line shows* $\widehat{\rho}$.

results from different sub-samples we have used proportional weights and size-proportional weights:

$$\begin{cases} \text{Prop} = \frac{C_k}{\sum_l C_l}, \\ \text{Size.Prop} = \frac{C_k n_k}{\sum_l C_l n_l}. \end{cases} . \tag{A.69}$$

The results are compared with full likelihood (Table A.5). In contrast to the compound-symmetry case, the size-proportional weights show much better results than the proportional weights. Furthermore, the size-proportional weights in the current simulation are identical with the equal weights. The n_k's have a much larger influence in the AR(1) case compared to CS. Figures A.2, A.3, and A.4 make the comparisons easier.

C.2 Simulations with Proportional and Size-proportional Weights: ρ near 0/1

We now present a comparison between proportional and size-proportional weights. We see that, for ρ's near 1 (i.e., near CS), size-proportional weights are worse than proportional weights.

We consider $c_1 = 500$, $c_2 = 250$, $c_3 = 250$, $c_4 = 500$, and $n_1 = 5$, $n_2 = 10$, $n_3 = 10$, $n_4 = 5$. Parameters are set as $\mu = 0$, $\sigma = 2$ and $\rho \in \{0.01, 0.2, 0.5, 0.8, 0.9, 0.95, 0.99\}$. The data are generated 100 times and the model is fitted using PROC MIXED in SAS (for a single overall intercept). For combining results from different sub-samples we have used proportional weights and size-proportional weights as in (A.69). The results are compared with full likelihood results.

Table A.5: *Comparing proportional, size-proportional and iterated optimal weights with full likelihood for AR(1) covariance structure.*

		$\widehat{\mu}$	Sd	$\widehat{\rho}$	Sd	$\widehat{\sigma}^2$	Sd
$\rho = 0.1$	Prop.	-0.00190	0.01615	0.10027	0.01158	1.99642	0.03002
	Size.Prop.	-0.00207	0.01538	0.10024	0.01080	1.99793	0.02853
	It.Opt.	-0.00206	0.01538	0.10024	0.00993	1.99792	0.02850
	ML	-0.00207	0.01538	0.10032	0.01078	1.99793	0.02850
$\rho = 0.5$	Prop.	-0.00212	0.02221	0.49966	0.00954	2.00349	0.03652
	Size.Prop	-0.00155	0.02156	0.49955	0.00898	2.00257	0.03494
	It.Opt.	-0.00168	0.02149	0.49956	0.00826	2.00265	0.03486
	ML	-0.00170	0.02150	0.49986	0.00896	2.00259	0.03488
$\rho = 0.8$	Prop.	0.00195	0.02890	0.79923	0.00549	1.99529	0.04989
	Size.Prop	0.00234	0.02904	0.79911	0.00530	1.99542	0.04907
	It.Opt.	0.00213	0.02855	0.79915	0.00486	1.99538	0.04859
	ML	0.00212	0.02855	0.79937	0.00527	1.99519	0.04861

In Figure A.5, for $\rho = 0.99$ and 0.95, the size-proportional weights perform worse than the proportional weights. This is expected, because in this case AR(1) approaches CS. This result is clearer in the left panel of Figure A.5, where the standard deviations are shown. For ρ's near 1, the proportional weights are as efficient as full likelihood, while as ρ moves further from 1 this would happen for size-proportional weights.

Figure A.6 shows this phenomenon more clearly, as for some selected ρ's $(0.01, 0.5, 0.95)$ the density plot for all 100 simulated datasets is plotted rather than a boxplot. The size-proportional weights are better than proportional weights if ρ is not very close to 1. As soon as ρ becomes 0.95 or 0.99, the size-proportional weights become worse.

C.3 Simulations With Optimal Weights

Given the covariance matrix of the parameter estimators, finding the optimal weights is straightforward, but in practice the unknown parameters therein need to be estimated. Here we compare the iterative weights with size-proportional weights and ML. See Figures A.7, A.8, A.9, and Table A.5. As expected, the optimal weights lead to estimates very close to the MLE; the difference between them being numerical.

Size-proportional weights are used as the initial weights to begin the iterative procedure. One interesting outcome of this simulation is that the iterative procedure always converged after just one iteration. This means, iterated optimal weights are just like the approximated optimal weights, but there, instead of using $\widehat{\theta}_k$ from each sub-sample, one may use $\tilde{\theta}$ obtained from all sub-samples using a non-optimal but good weight.

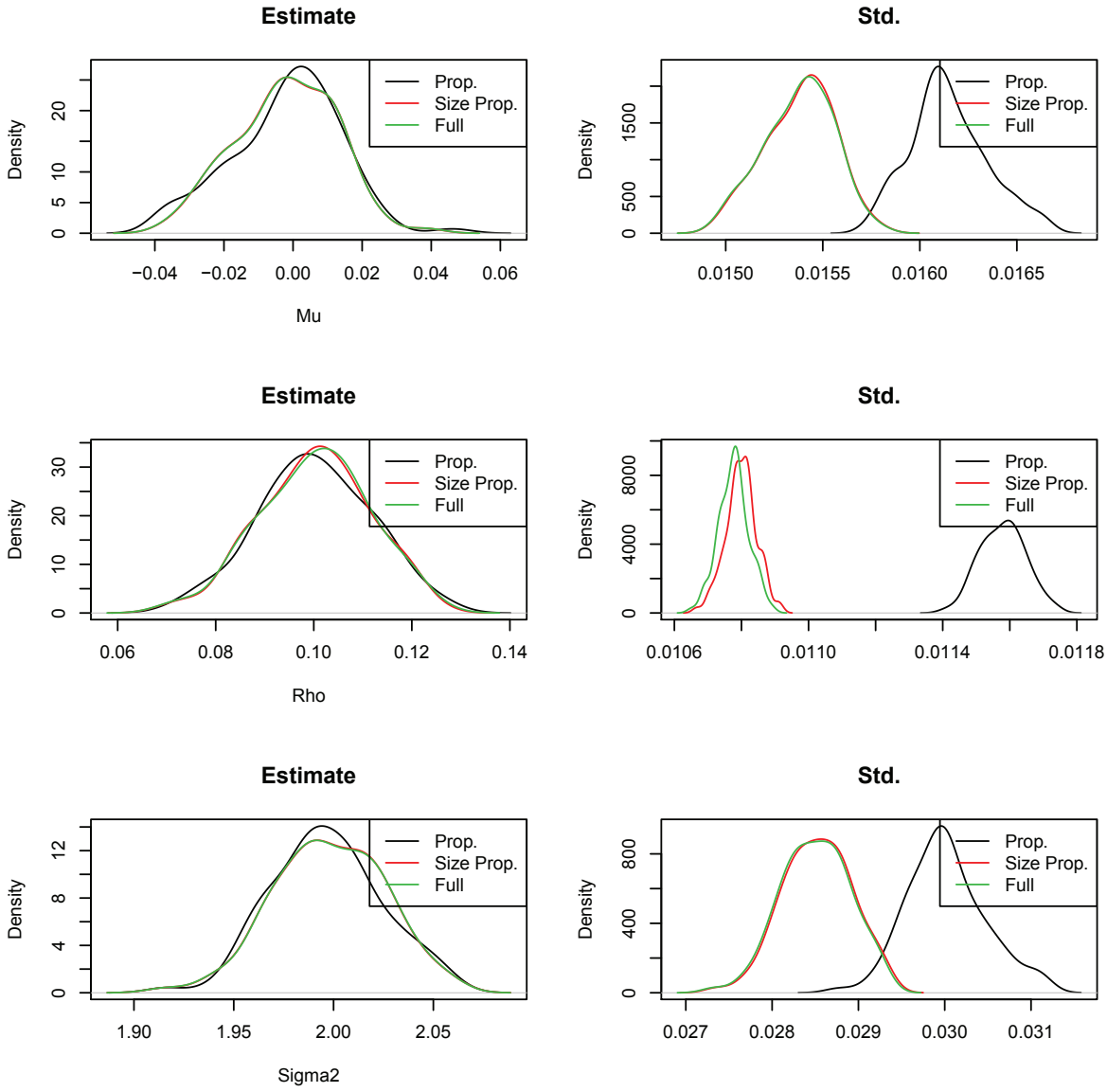

Figure A.2: *Comparing proportional and size-proportional weights with full likelihood for* 100 *replications with* $\mu = 0$, $\sigma^2 = 2$ *and* $\rho = 0.1$.

C.4 Simulations on Computation Time

Here, some summary tables are presented to summarize the results which are already presented via figures earlier. Furthermore, a table and a figure are added to compare computation time for closed form solutions to numerical ones.

In each table the mean of the estimated parameter and its standard deviation using the 100 replications are given, together with the standard deviation of those 100 numbers (in parentheses). If θ is the parameter of interest, $\widehat{\theta}$ is its estimate and θ_0 is its real value, then the MSE is computed

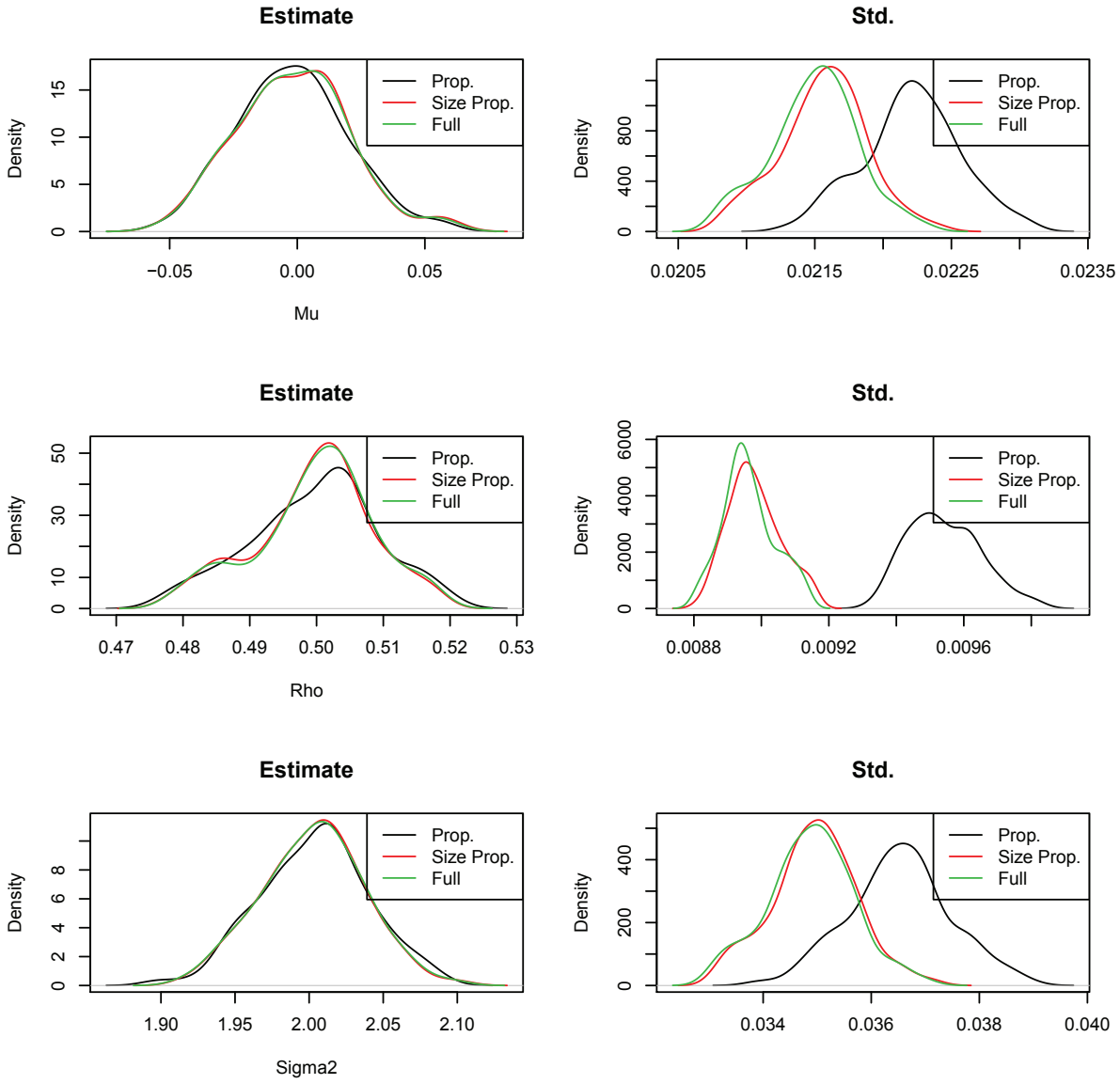

Figure A.3: *Comparing proportional and size-proportional weights with full likelihood for* 100 *replications with* $\mu = 0$, $\sigma^2 = 2$ *and* $\rho = 0.5$.

as follows:

$$\text{MSE}(\widehat{\theta}) = \frac{1}{100}\sum_{i=1}^{100}(\widehat{\theta} - \theta_0)^2. \tag{A.70}$$

Table A.6 summarizes the results for μ. The sample splitting estimates are computed using proportional and size-proportional (identical to equal weights in this case) weights. The results using the full sample are also given. The third column in Table A.6 presents the averaged (over 100 replications) estimated μ and its standard deviation. The fourth column presents the averaged estimated standard deviation for $\widehat{\mu}$ (over 100 replications) and its standard deviation. The last column shows the MSE computed using (A.70) for $\mu_0 = 1$. Tables A.7 and A.8 shows the same results for ρ and

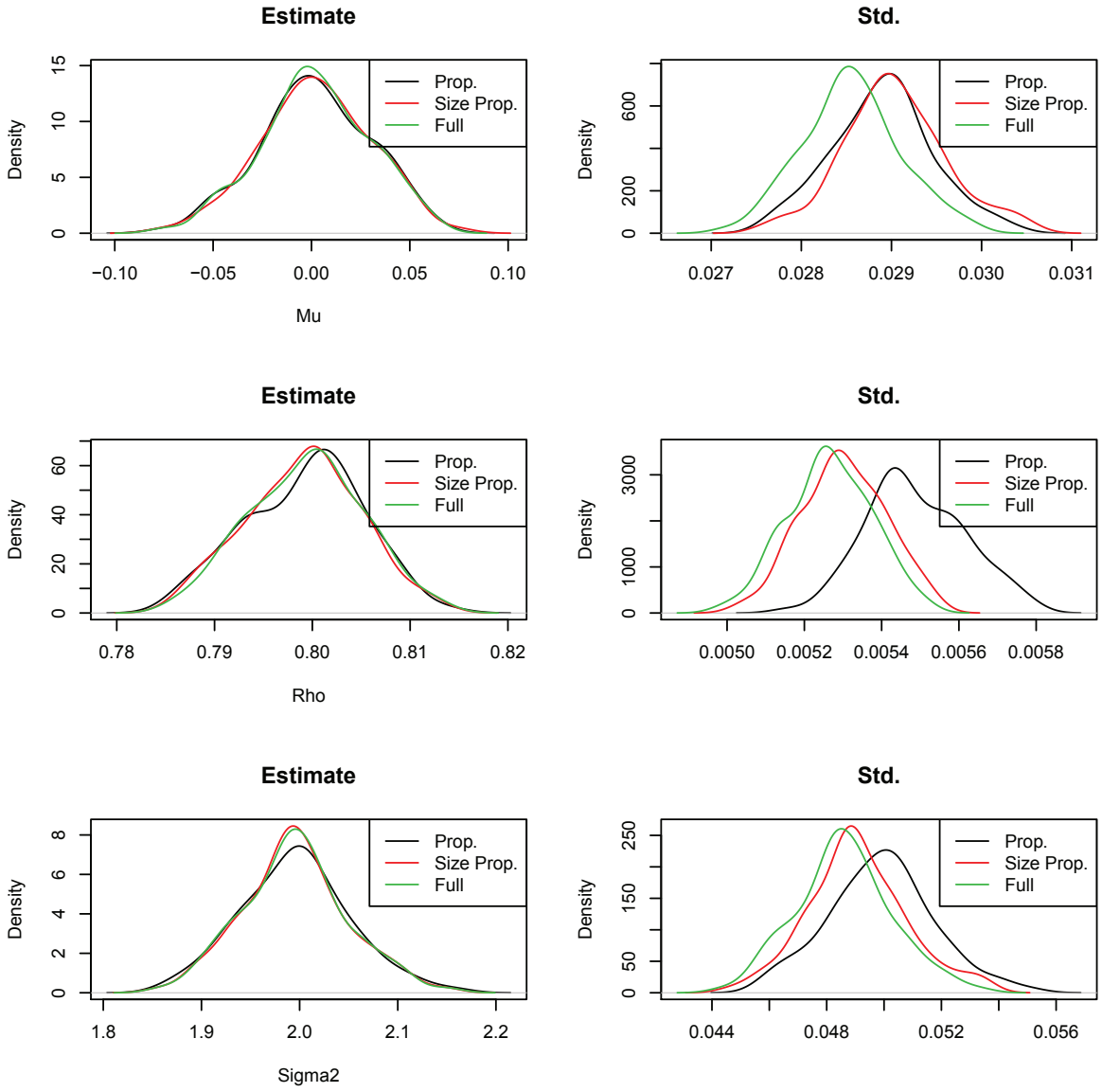

Figure A.4: *Comparing proportional and size-proportional weights with full likelihood for* 100 *replications with* $\mu = 0$, $\sigma^2 = 2$ *and* $\rho = 0.8$.

σ^2 ($\sigma_0^2 = 2$), respectively.

Table A.9 compares the computation time between closed-form and interative methods. The closed-form solutions are implemented in R and for the numerical methods the MIXED procedure in SAS is used, with error covariance structure set to AR(1).

The data are generated using $n = 10$ for all clusters, with c is varying from 100 to 1000000. Therefore, the design is balanced and the point of this comparison is to see how the computation time is reduced in each split.

As one may see in Table A.9 and Figure A.10, using closed form solutions significantly reduces the computation time. This means, as well as the computation time reduction due to splitting the

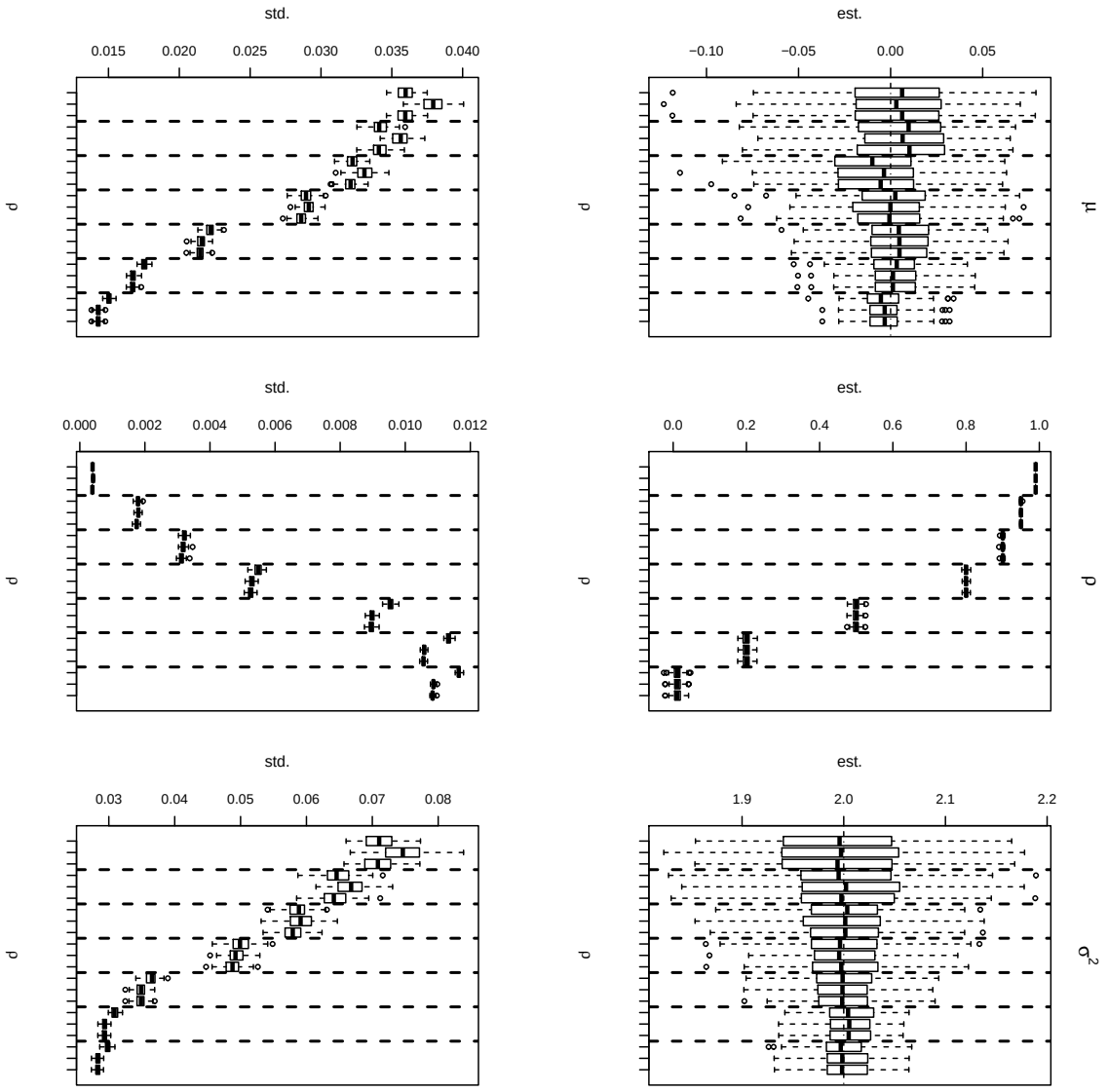

Figure A.5: *Simulation study. Boxplots comparing proportional and size-proportional weights with full likelihood for* 100 *replications with* $\mu = 0$, $\sigma^2 = 2$ *and* $\rho = 0.99, 0.95, 0.9, 0.8, 0.5, 0.2, 0.01$. *In every section of the boxplots (which are separated by dashed lines) the first out of three represents the proportional weights, the middle of is size-propotional weights and the one on the right shows the results for the full likelihood. The first row presents the estimates while the second row shows the standard deviations of these estimates.*

data, using closed form solutions within each split the computation time reduction is also huge: for example, for a million clusters, the reduction is from almost one hour to less than 5 seconds. Figure A.10 shows that computation time using closed form solution changes linearly with the number of clusters, while this will be exponential using an iterative solution.

To assess the effect of the overall size of the dataset, the model is fitted to two concatenated copies of the same set. Computation time results are presented in Table A.9 and Figure A.10. The data are generated with $\mu = 0$, $\sigma^2 = 2$, and $\rho = 0.25$.

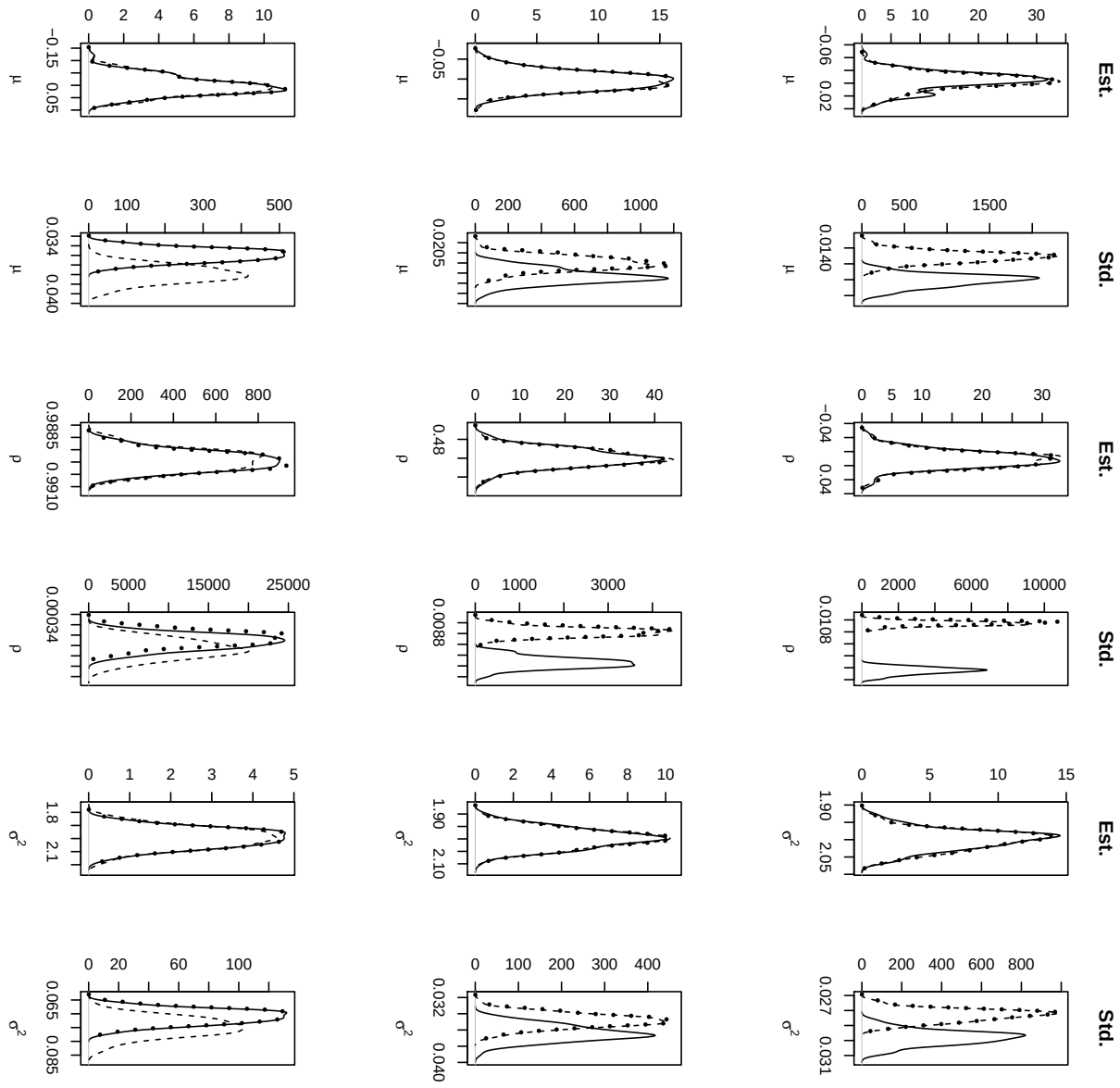

Figure A.6: *Simulation study. Comparing proportional, size-proportional and full likelihood results via their empirical density for the 100 replications. In all of the figures* $\mu = 0$ *and* $\sigma^2 = 2$*. The first row is for* $\rho = 0.01$*, the middle one is for* $\rho = 0.5$ *and last one corresponds to* $\rho = 0.99$*. In each figure the ticker dotted line corresponds to full likelihood, the dashed line is for size-proportional weights and the solid line is for proportional weights.*

D Details on PANSS Data Analysis

As one may see from Table A.10, by far the majority of the study subjects have complete data and hence belong to the first pattern.

Figure A.11 presents boxplots for the entire set of data, for the subjects from the first pattern only, and for various split samples.

To examine the choice of an AR(1) covariance structure, Table A.11 shows three model selection

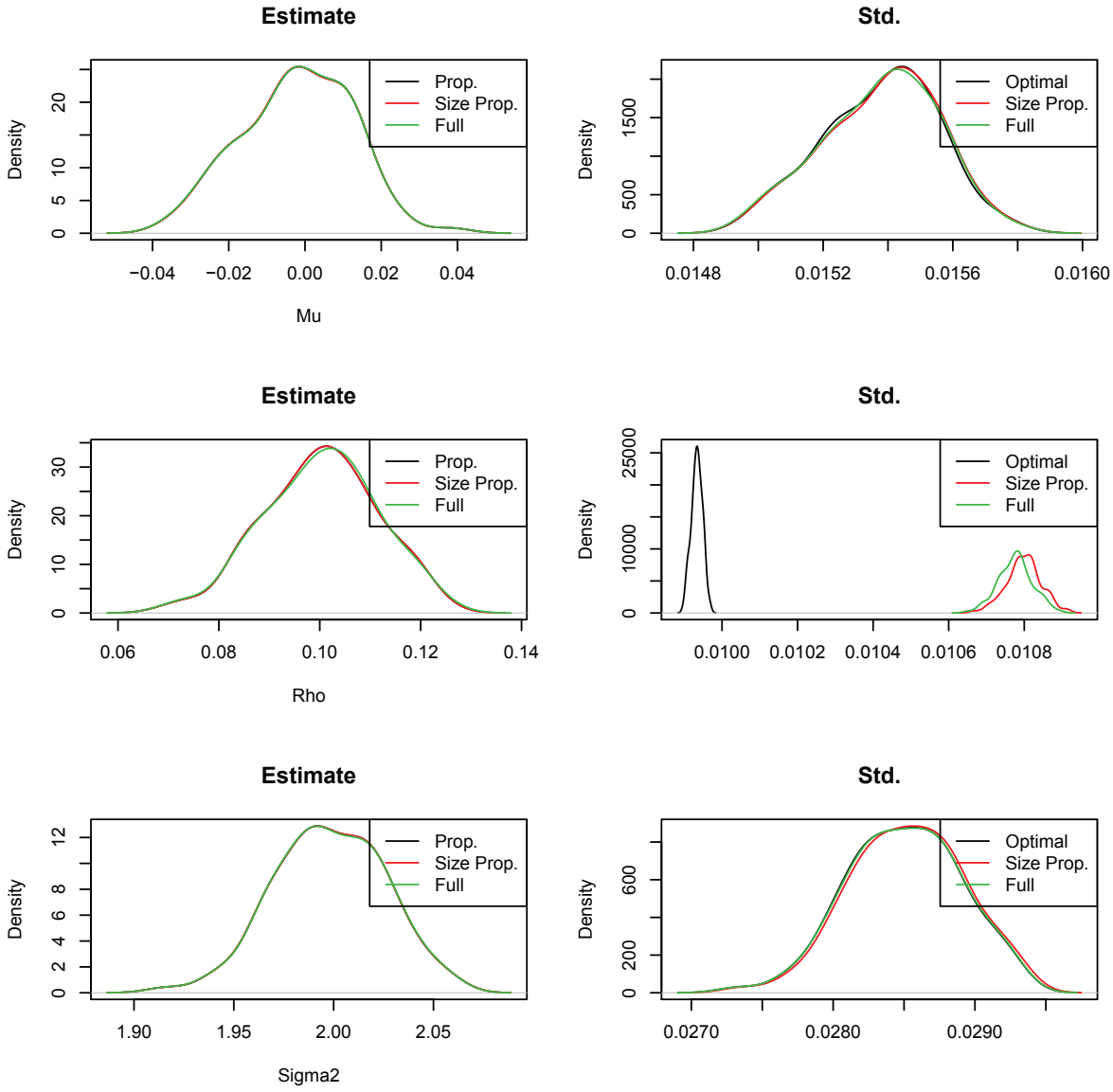

Figure A.7: *Comparing iterated optimal and size-proportioanl weights with full likelihood for* 100 *replications with* $\mu = 0$, $\sigma^2 = 2$ *and* $\rho = 0.1$.

criteria to compare different error covariance structures. Changing from independence structure ($R = \sigma^2 I$) to compound-symmetry ($R = \sigma^2 I$) the criteria decrease with a large amount, and the same when changing to AR(1). The step to an unstructured covariance does not make a big difference (considering that the unstructured covariance would has 21 parameters to estimate compared to 2 parameters in the AR(1) model). Therefore, AR(1) seems to be a good choice.

The 95% confidence intervals, accompanying (26), are presented in Figure A.12. In order to give more insight in these results, Figure A.13 shows the 95% confidence interval in each split, comparing with the full sample splits (the horizontal dashed line in the figure).

The 95% confidence intervals, accompanying (27), are presented in Figure A.14. Figure A.15

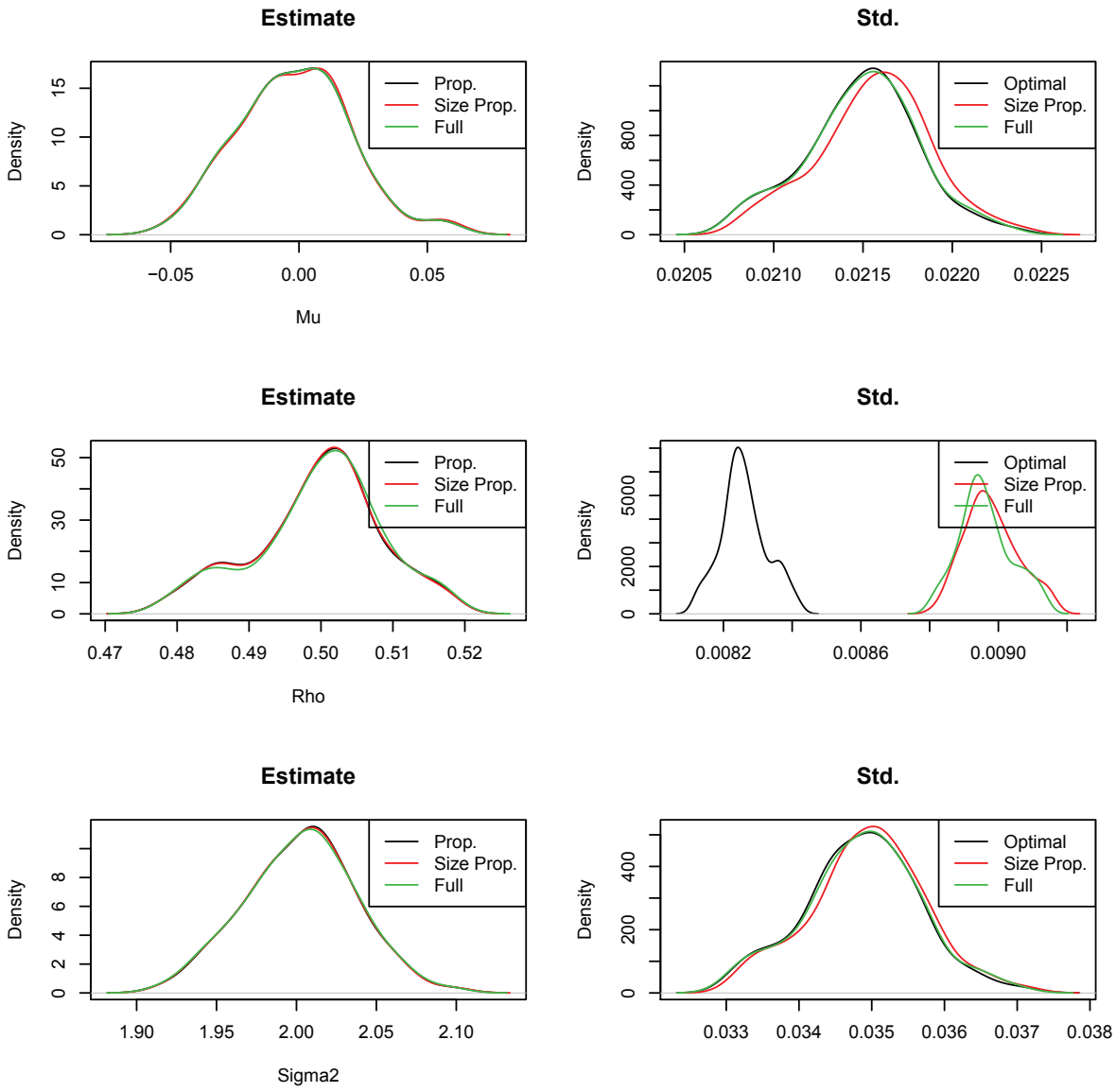

Figure A.8: *Comparing iterated optimal and size-proportioanl weights with full likelihood for* 100 *replications with* $\mu = 0$, $\sigma^2 = 2$ *and* $\rho = 0.5$.

shows the 95% confidence intervals for the parameter estimates in each split comparing with the full sample estimate (the horizontal dashed-like in the figure.)

E R Code

Estimating variance components

```
est.ar1 <- function(C,n,Y,Plot=1){


# making a matrix out of the response vector
```

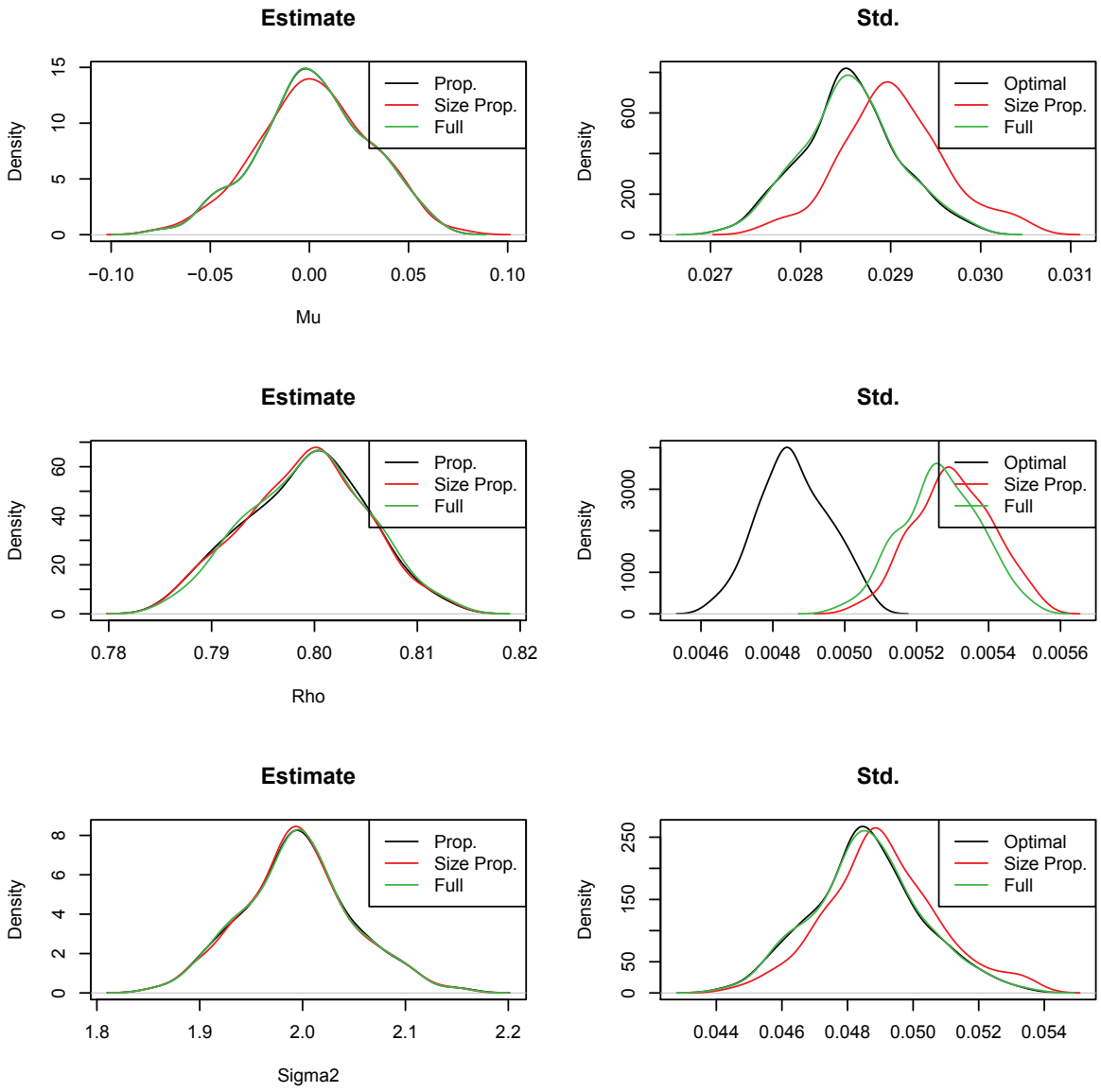

Figure A.9: *Comparing iterated optimal and size-proportioanl weights with full likelihood for* 100 *replications with* $\mu = 0$, $\sigma^2 = 2$ *and* $\rho = 0.8$.

```
Resp=matrix(Y,n,C)


# Computing cross products

SS=crossprod(t(Resp))


# Computing S, \tilde{S} and R

S=sum(diag(SS))

S.tilde=sum(diag(SS)[2:(n-1)])

tmp.R=SS
```

Table A.6: *Simulation study. Estimating μ and its standard deviation. The mean (standard deviation) of the* 100 *replications are given together with mean squared errors for* $\rho = 0.01, 0.2, 0.5, 0.8, 0.9, 0.99$ *using proportional and size-proportional weights comparing with the full likelihood results.*

ρ_0	method	mean$(\widehat{\mu})$ (s.d.)	mean(s.e.$(\widehat{\mu})$) (s.d.)	MSE$\times 10^4$
	Prop.	-0.00271 (0.01462)	0.01503 (0.00020)	2.19000
0.01	Size Prop.	-0.00277 (0.01320)	0.01429 (0.00018)	1.80172
	Full	-0.00275 (0.01319)	0.01428 (0.00018)	1.79828
	Prop.	0.00158 (0.01677)	0.01752 (0.00025)	2.81056
0.2	Size Prop.	0.00085 (0.01616)	0.01673 (0.00021)	2.59147
	Full	0.00090 (0.01615)	0.01672 (0.00021)	2.58880
	Prop.	0.00391 (0.02244)	0.02217 (0.00037)	5.13770
0.5	Size Prop.	0.00397 (0.02191)	0.02155 (0.00035)	4.91201
	Full	0.00396 (0.02182)	0.02148 (0.00034)	4.87038
	Prop.	0.00130 (0.02790)	0.02894 (0.00050)	7.72450
0.8	Size Prop.	-0.00049 (0.02710)	0.02912 (0.00045)	7.27464
	Full	0.00053 (0.02713)	0.02862 (0.00045)	7.29130
	Prop.	-0.00828 (0.03006)	0.03221 (0.00056)	9.63393
0.9	Size Prop.	-0.00727 (0.03145)	0.03306 (0.00070)	10.3224
	Full	-0.00803 (0.02998)	0.03207 (0.00057)	9.54258
	Prop.	0.00162 (0.03663)	0.03597 (0.00064)	13.3123e
0.99	Size Prop.	-0.00007 (0.03930)	0.03797 (0.00088)	15.2876
	Full	0.00156 (0.03666)	0.03597 (0.00064)	13.3305

```
diag(tmp.R)=NA

tmp.R2 = (matrix(tmp.R[which(!is.na(tmp.R))],nrow=n,ncol=n-1))

R=sum(tmp.R2[1,])


# Finding the coefficients of the 3rd degree polynomial and its roots

P1=(n-1)*S.tilde

P2=(n-2)*R

P3=((n*S.tilde)+S)

P4=n*R


PP=polynomial(c(P4,-P3,-P2,P1))

roots=polyroot(PP)

Roots=Re(roots)[abs(Im(roots)) < 1e-6]

rho.hat=Roots[abs(Roots)<1]
```

Table A.7: *Simulation study. Estimating ρ and its standard deviation. The mean (standard deviation) of the* 100 *replications are given together with mean squared errors for* $\rho = 0.01, 0.2, 0.5, 0.8, 0.9, 0.99$ *using proportional and size-proportional weights comparing with the full likelihood results.*

ρ_0	method	mean($\widehat{\rho}$) (s.e.)	mean(s.e.($\widehat{\rho}$)) (s.e.)	MSE
	Prop.	0.01077 (0.01237)	0.01165 (0.00006)	1.52178e-04
0.01	Size Prop.	0.01115 (0.01200)	0.01087 (0.00004)	1.43974e-04
	Full	0.01123 (0.01203)	0.01084 (0.00004)	1.44675e-04
	Prop.	0.19960 (0.01213)	0.01133 (0.00007)	1.45806e-04
0.2	Size Prop.	0.19973 (0.01145)	0.01058 (0.00005)	1.29974e-04
	Full	0.19986 (0.01142)	0.01056 (0.00005)	1.29174e-04
	Prop.	0.49956 (0.00963)	0.00954 (0.00011)	9.19119e-05
0.5	Size Prop.	0.49965 (0.00904)	0.00898 (0.00008)	8.11057e-05
	Full	0.49990 (0.00904)	0.00896 (0.00008)	8.08877e-05
	Prop.	0.79973 (0.00541)	0.00548 (0.00012)	2.90660e-05
0.8	Size Prop	0.79990 (0.00483)	0.00529 (0.00009)	2.31132e-05
	Full	0.80018 (0.00489)	0.00525 (0.00009)	2.36726e-05
	Prop.	0.90017 (0.00286)	0.00321 (0.00008)	8.14862e-06
0.9	Size Prop.	0.90013 (0.00297)	0.00318 (0.00008)	8.76257e-06
	Full	0.90040 (0.00292)	0.00312 (0.00008)	8.60114e-06
	Prop.	0.98994 (0.00038)	0.00039 (0.00001)	1.45292e-07
0.99	Size Prop.	0.98992 (0.00042)	0.00041 (0.00002)	1.77848e-07
	Full	0.98997 (0.00037)	0.00039 (0.00001)	1.37289e-07

```
# Plotting the 3rd degree polynomial if requested


if (Plot==1){

plot(PP,xlim=c(-1.5,1.5),xlab="rho",ylab="3rd degree polynomial")

abline(h=0,col=2)

abline(v=Roots[abs(Roots)<1],lty=2,lwd=2,col=2)

abline(v=-1,lty=2)

abline(v=1,lty=2)

}


# Estimating \sigma2


tmp1=1/(C*n)

tmp2=1/(1-(rho.hat^2))

tmp3=S+((rho.hat^2)*S.tilde)
```

Table A.8: *Simulation study. Estimating* σ^2 *and its standard deviation. The mean (standard deviation) of the* 100 *replications are given together with mean squared errors for* $\rho = 0.01, 0.2, 0.5, 0.8, 0.9, 0.99$ *using proportional and size-proportional weights comparing with the full likelihood results.*

ρ_0	method	mean($\widehat{\sigma}^2$) (s.e.)	mean(s.e.($\widehat{\sigma}^2$)) (s.e.)	MSE
	Prop.	1.99964 (0.02960)	0.02981 (0.00049)	8.67280e-04
0.01	Size Prop.	2.00165 (0.02836)	0.02834 (0.00040)	7.98842e-04
	Full	2.00167 (0.02832)	0.02832 (0.00040)	7.97002e-04
	Prop.	2.00581 (0.02907)	0.03093 (0.00055)	8.70077e-04
0.2	Size Prop.	2.00484 (0.02778)	0.02936 (0.00044)	7.87298e-04
	Full	2.00473 (0.02772)	0.02933 (0.00044)	7.83248e-04
	Prop.	1.99783 (0.03860)	0.03638 (0.00097)	1.47960e-03
0.5	Size Prop.	1.99890 (0.03747)	0.03488 (0.00085)	1.39116e-03
	Full	1.99897 (0.03748)	0.03481 (0.00085)	1.39152e-03
	Prop.	2.00013 (0.04900)	0.05002 (0.00166)	2.37744e-03
0.8	Size Prop.	2.00136 (0.04423)	0.04930 (0.00137)	1.93881e-03
	Full	2.00101 (0.04569)	0.04880 (0.00142)	2.06754e-03
	Prop.	2.00122 (0.05767)	0.05872 (0.00196)	3.29407e-03
0.9	Size Prop.	2.00089 (0.06037)	0.05915 (0.00230)	3.60829e-03
	Full	2.00115 (0.05876)	0.05793 (0.00198)	3.41986e-03
	Prop.	1.99683 (0.06941)	0.07117 (0.00254)	4.77911e-03
0.99	Size Prop.	1.99527 (0.07598)	0.07484 (0.00344)	5.73813e-03
	Full	1.99641 (0.06940)	0.07093 (0.00253)	4.78099e-03

```
tmp4=2*rho.hat*R


sigma2.hat=(tmp1*tmp2)*(tmp3-tmp4)


return(list(rho.hat=rho.hat,sigma2.hat=sigma2.hat))
}
```

Computing variance of parameter estimates

```
cov.ar1 <- function(ck,nk,rho,sigma2){
num.split=length(ck)
var.mu1=(ck/(sigma2*(1-(rho^2))))* (((nk-2)*rho^2)-(2*((nk-1)*rho))+nk)
var.mu=1/var.mu1
w.mu=var.mu1/sum(var.mu1)


# Note that the unbiased version of the covariance is used here
```

Table A.9: *Simulation study. The computation time for a sample with* $n = 10$ *and* c *=1e+02, 1e+03, 1e+04, 5e+04, 1e+05, 3e+05, 5e+05, 7e+05, 9e+05, 1e+06. The closed form solution is obtained by implementing the results of this paper in R, and the numerical solution is obtained using PROC MIXED in SAS to estimate a repeated measurement model with AR(1) covariance structure.*

time (s)	1e+02	1e+03	1e+04	5e+04	1e+05	3e+05	5e+05	7e+05	9e+05	1e+06
Closed form	0.00	0.00	0.03	0.23	0.34	1.45	2.07	3.37	4.40	4.89
Numerical	0.08	0.13	1.04	10.45	34.74	268.96	770.74	1611.43	2724.31	3399.47

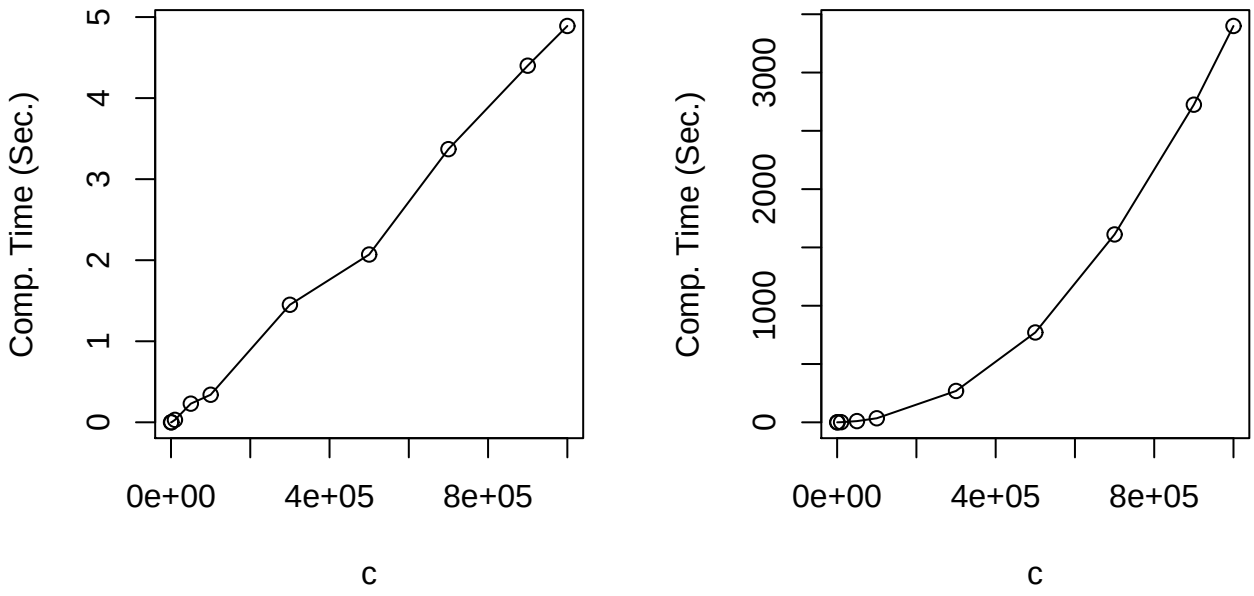

Figure A.10: *Simulation study. Comparing computation time using closed form (left) and numerical (right) solutions. The horizontal axis shows number of clusters (c) and the vertical axis shows the computation time in seconds.*

```
v22=2*(sigma2^2)*(1+(rho^2))

v12=2*sigma2*(1-(rho^2))

v11=(1-rho^2)^2

var.varcomp1=matrix(c(v11,v12,v12,v22),2,2)

varcomp.coef=1/(ck*(nk-((nk-2)*(rho^2))))

var.varcomp=outer(var.varcomp1,varcomp.coef)


W.total=0

for (i in 1:num.split){
```

Table A.10: *PANSS data. Number of clusters in each trial for each cluster pattern.*

											Trial					
n	Pattern										FIN-1	FRA-3	INT-2	INT-3	INT-7	Total
	∗	∗	·	·	·	·	·	·	·	·	17	8	71	43	3	142
2	∗	·	∗	·	·	·	·	·	·	·	0	0	2	0	1	3
	∗	·	·	·	∗	·	·	·	·	·	0	0	1	0	0	1
	∗	∗	∗	·	·	·	·	·	·	·	8	4	83	41	7	143
3	∗	·	∗	·	∗	·	·	·	·	·	0	0	2	0	0	2
	∗	∗	·	·	∗	·	·	·	·	·	1	0	3	1	0	5
	∗	∗	∗	·	∗	·	·	·	·	·	11	0	85	66	5	167
	∗	·	∗	·	∗	·	∗	·	·	·	0	0	1	0	1	2
4	∗	·	∗	·	∗	·	·	∗	·	·	0	0	1	0	0	1
	∗	∗	∗	·	·	·	∗	·	·	·	0	0	3	0	0	3
	∗	∗	∗	∗	·	·	·	·	·	·	0	4	1	0	0	5
	∗	∗	·	∗	·	·	∗	·	·	·	0	1	0	0	0	1
	∗	·	∗	·	·	·	∗	∗	·	·	0	0	0	0	1	1
	∗	∗	∗	·	∗	·	∗	·	·	·	58	0	85	35	6	184
	∗	∗	∗	·	∗	·	·	∗	·	·	0	0	8	0	1	9
	∗	∗	·	·	∗	·	∗	∗	·	·	0	0	6	0	0	6
5	∗	∗	∗	·	·	·	∗	∗	·	·	0	0	8	0	0	8
	∗	·	∗	·	∗	·	∗	∗	·	·	0	0	3	0	2	5
	∗	·	∗	·	∗	·	·	∗	·	∗	0	0	1	0	0	1
	∗	∗	∗	∗	∗	·	·	·	·	·	0	44	0	0	0	44
	∗	∗	·	∗	∗	∗	·	·	·	·	0	1	0	0	0	1
	∗	∗	∗	·	∗	·	∗	∗	·	·	0	0	986	240	74	1300
	∗	∗	·	·	∗	·	∗	∗	∗	·	0	0	1	0	0	1
6	∗	∗	∗	·	·	·	∗	∗	·	∗	0	0	1	0	0	1
	∗	∗	∗	·	∗	·	∗	·	∗	·	0	0	1	0	0	1
	∗	∗	∗	·	∗	∗	∗	·	·	·	0	0	2	0	0	2

```
W.total=W.total+solve(var.varcomp[,,i])
}
w.varcomp=array(0,c(2,2,num.split))
for (i in 1:num.split){
w.varcomp[,,i]=solve(W.total)%*%solve(var.varcomp[,,i])
}

return(list(var.mu=var.mu,var.varcomp=var.varcomp
,w.mu=w.mu,w.varcomp=w.varcomp))
}
```

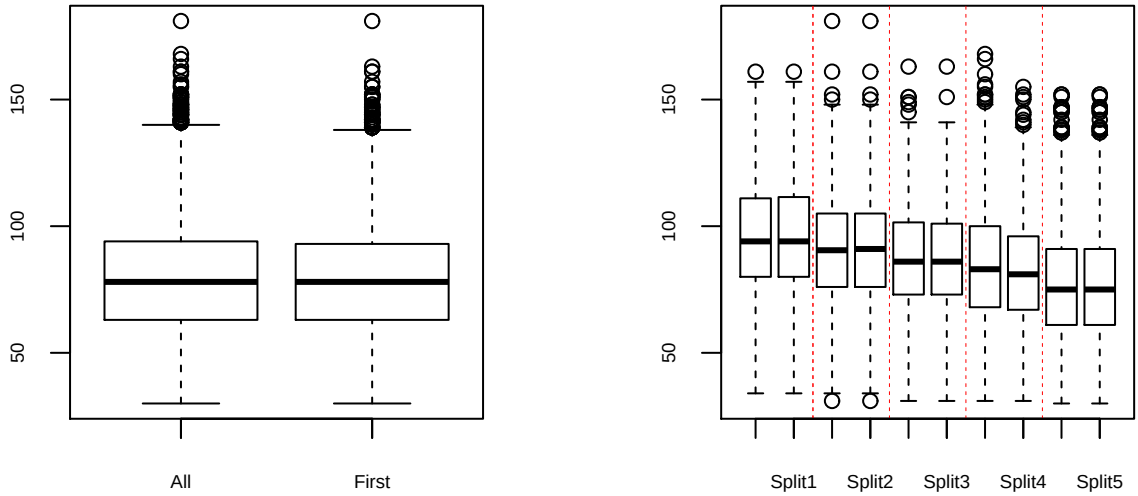

Figure A.11: *PANSS data. Boxplots for the entire set of data, for the subject from the first pattern only, and for various split samples.*

Table A.11: *PANSS data. Comparing different error covariance structures using three model comparison criteria for model (26) (residual log-likelihood value; AIC; BIC). Three R structures: Ind. : independence structure ($R = \sigma^2 I$), CS: compound-symmetry structure ($R = \sigma^2 I + dJ$), AR(1): AR(1) structure ($R_{ij} = \sigma^2 \rho^{|i-j|}$), UN: unstructured ($R_{ij} = \sigma_{ij}^2$).*

Model	-2 Res.log.lik.	AIC	BIC
Unstructured	80005.1	80047.1	80164.1
AR(1)	80522.6	80526.6	80537.8
Compound symm.	82683.1	82687.1	82698.3
Independence	89546.1	89548.1	89553.7

Computing iterated optimal weights

```
iterate.optimal.ar1 <- function(ck,nk,u.split,var.comp.split,tol){


num.split=length(ck)

W.size.prop=(nk*ck)/sum(nk*ck)

rho.hat= sum(var.comp.split[1,]*W.size.prop)

sigma2.hat=sum(var.comp.split[2,]*W.size.prop)


diff=10

var.comp.hat=matrix(c(rho.hat,sigma2.hat),2,1)

count=0
```

```
while (diff>tol){
WW=cov.ar1 (ck,nk,var.comp.hat[1,1],var.comp.hat[2,1])
W.mu=WW$w.mu
W.varcomp=WW$w.varcomp
var.comp.hat=0
for (i in 1:num.split){
var.comp.hat=var.comp.hat+W.varcomp[,,i]%*%var.comp.split[,i]
}
W.mu.old=W.mu
W.varcomp.old=W.varcomp
WW=cov.ar1 (ck,nk,var.comp.hat[1,1],var.comp.hat[2,1])
W.mu=WW$w.mu
W.varcomp=WW$w.varcomp
diff1=norm(as.matrix(W.mu-W.mu.old))
diff2=sum(apply(W.varcomp-W.varcomp.old,3,norm))
diff=max(c(diff1,diff2))
count=count+1
}
var.comp.hat=0
for (i in 1:num.split){
var.comp.hat=var.comp.hat+W.varcomp[,,i]%*%var.comp.split[,i]
}

W.total=0
WW=cov.ar1 (ck,nk,var.comp.hat[1,1],var.comp.hat[2,1])
for (i in 1:num.split){
W.total=W.total+solve(WW$var.varcomp[,,i])
}

mu.hat=sum(W.mu*mu.split)

return(list(W.mu=W.mu,W.varcomp=W.varcomp,mu.hat=mu.hat
```

```
,varcomp.hat=var.comp.hat, var.mu.hat=1/sum(1/WW$var.mu),
var.varcomp.hat=solve(W.total),num.iterate=count))
}
```

References

Franci, R., and Rigatelli, L.T. (1979). *Storia della teoria delle equazioni algebriche.* (Vol. 40). MIlan: Ugo Mursia Editore.

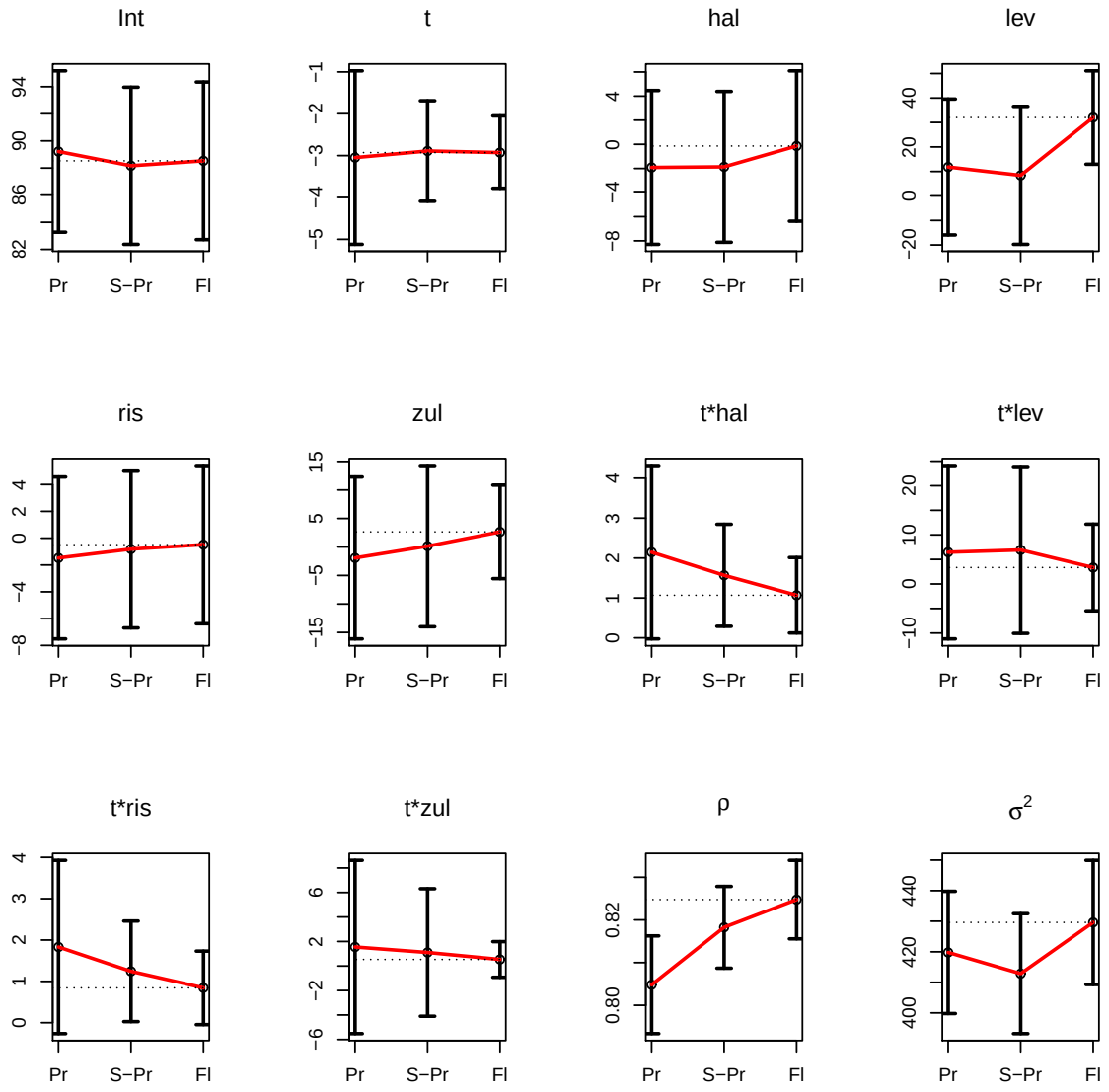

Figure A.12: *PANSS data.* 95% *confidence intervals for fixed effects and variance components estimates and the standard deviations of these estimates using sample splitting, combined with proportional (Pr - first) and size-proportional (S-Pr - second) weights, and full likelihood (Fl - third). The dashed horizontal line shows the full likelihood estimate. The model used in here is without trial effect (26).*

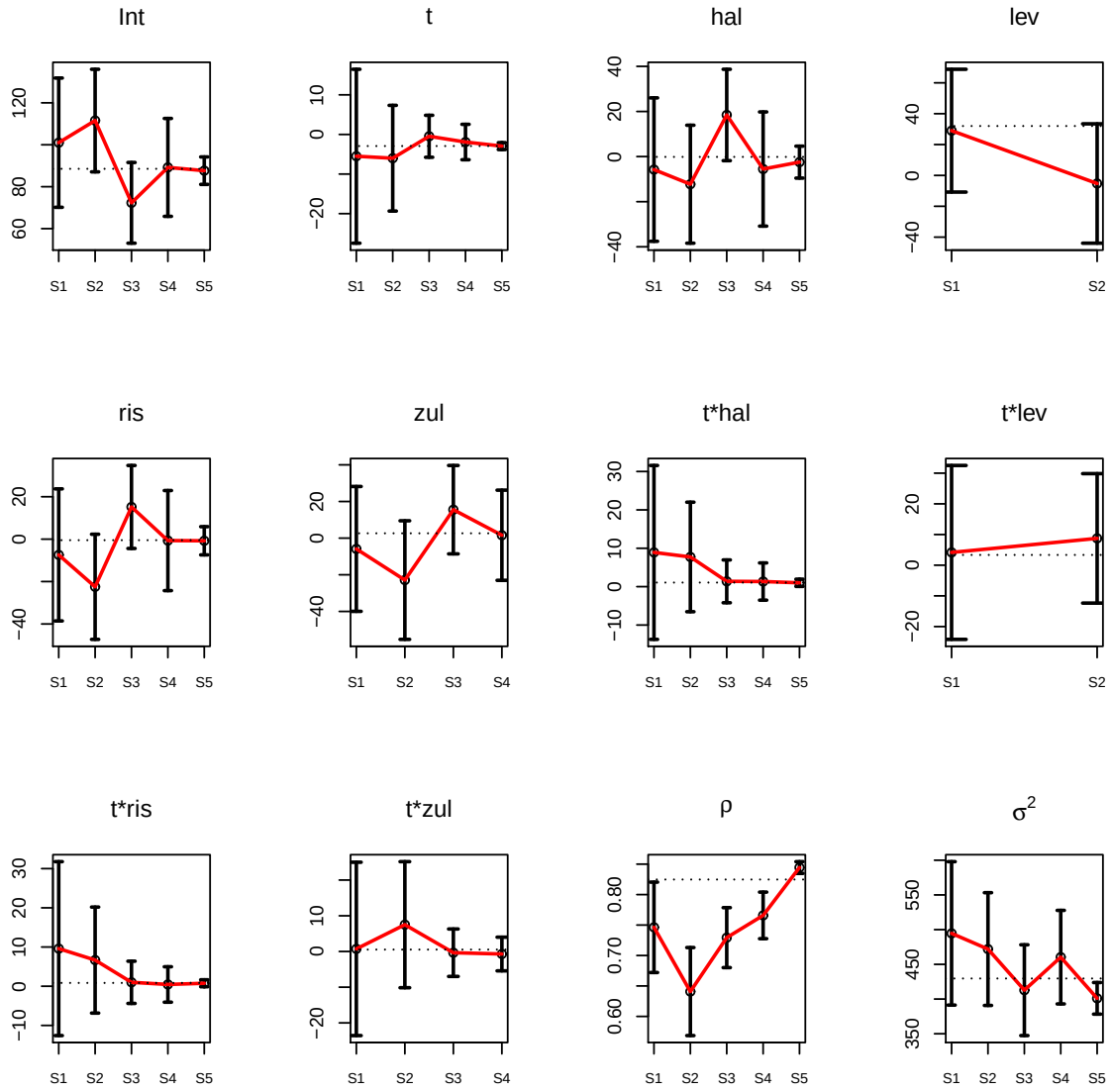

Figure A.13: *PANSS data.* 95% *confidence intervals for fixed effects and variance components estimates and the standard deviations of these estimates within each split. The dashed horizontal line shows the full likelihood estimate. The model used in here is without trial effect (27).*

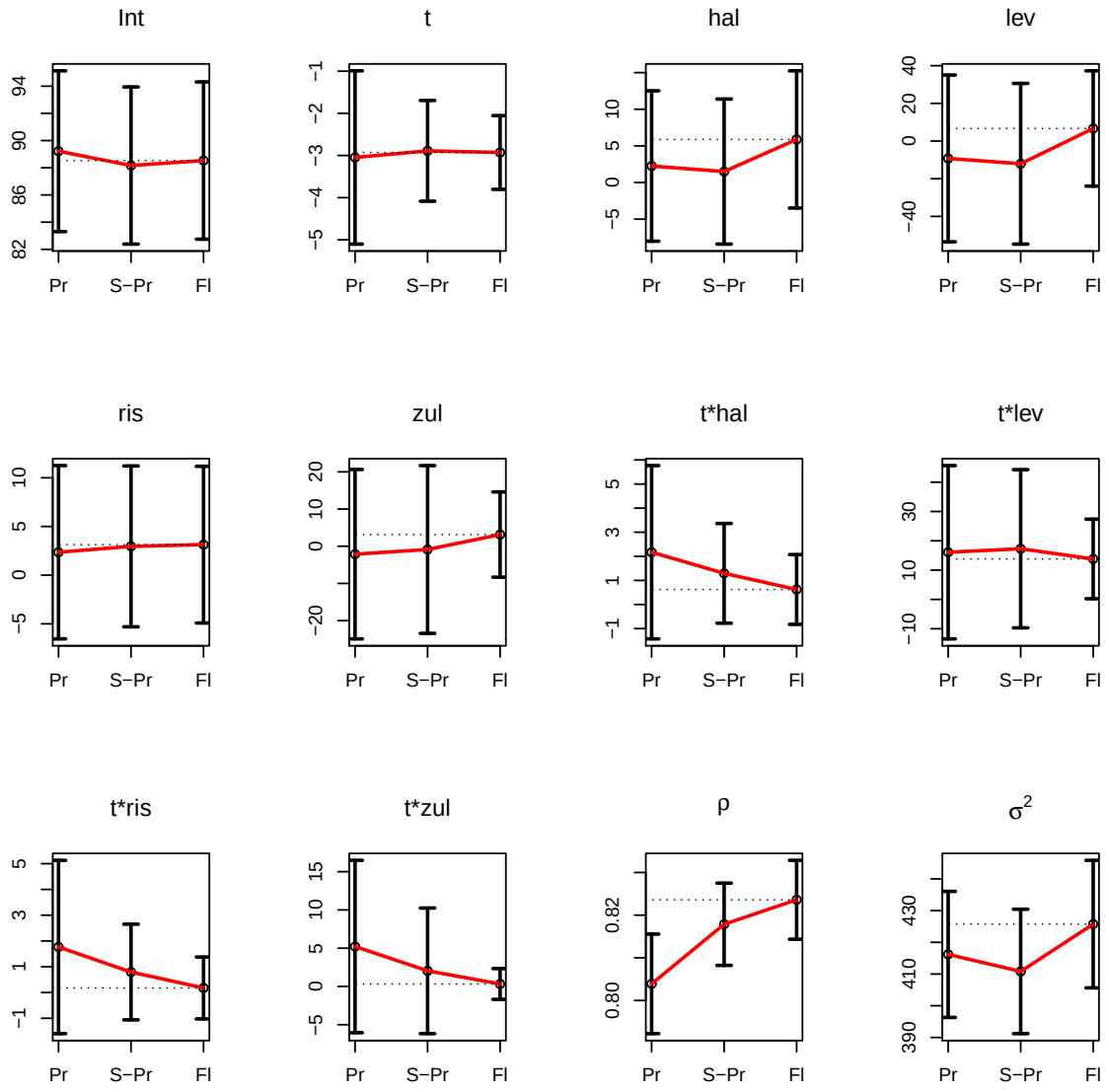

Figure A.14: *PANSS data.* 95% *confidence intervals for fixed effects and variance components estimates and the standard deviations of these estimates using sample splitting, combined with proportional (Pr - first) and size-proportional (S-Pr - second) weights, and full likelihood (Fl - third). The dashed horizontal line shows the full likelihood estimate. The model used in here is with trial effect (27).*

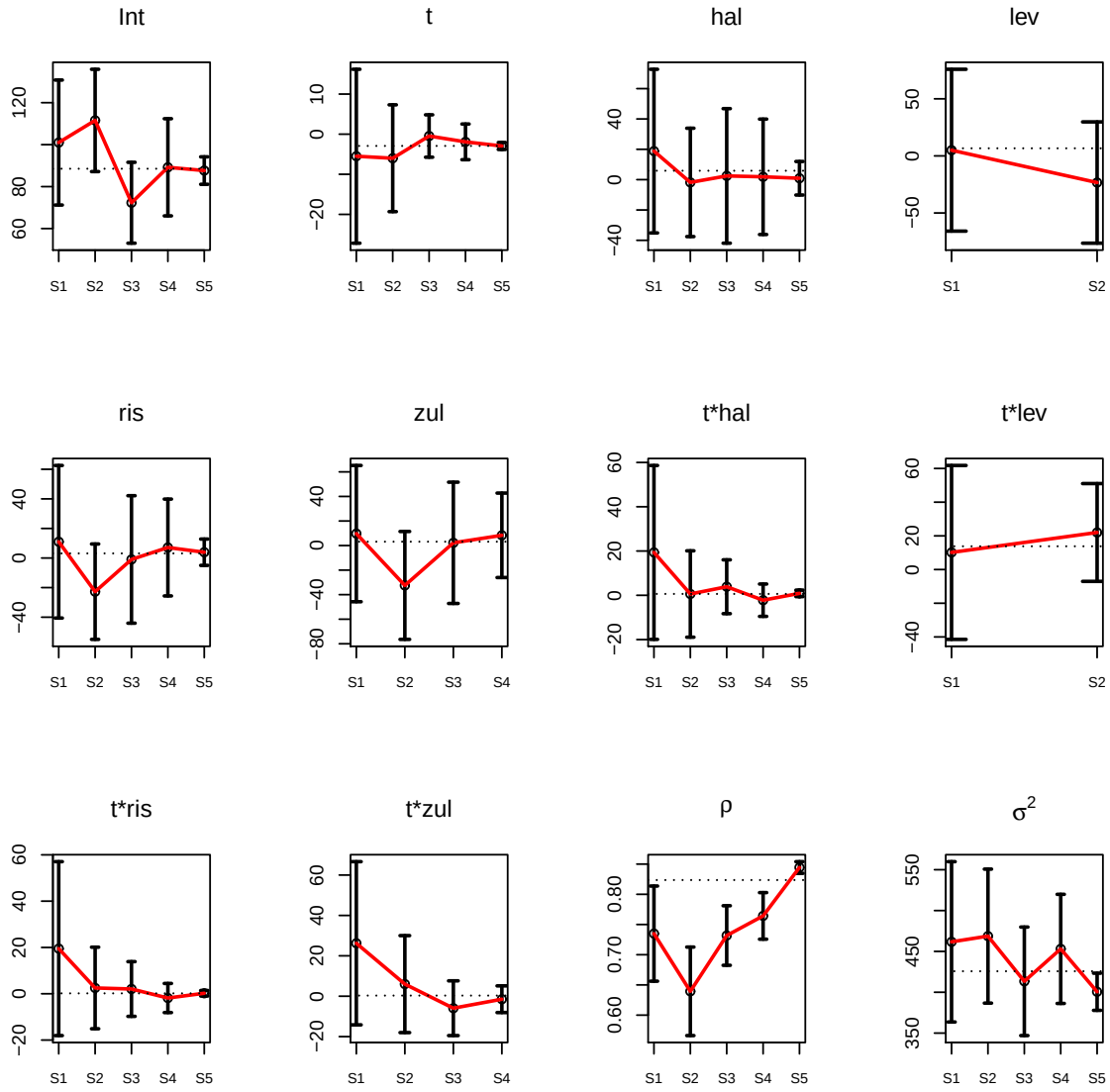

Figure A.15: *PANSS data.* 95% *confidence intervals for fixed effects and variance components estimates and the standard deviations of these estimates within each split. The dashed horizontal line shows the full likelihood estimate. The model used in here is with trial effect (27).*