Fast, closed-form, and efficient estimators for hierarchical models with AR(1) covariance and unequal cluster sizes

Lisa Hermans, Vahid Nassiri, Geert Molenberghs, Michael G. Kenward, Wim Van der Elst, Marc Aerts & Geert Verbeke

To cite this article: Lisa Hermans, Vahid Nassiri, Geert Molenberghs, Michael G. Kenward, Wim Van der Elst, Marc Aerts & Geert Verbeke (2017): Fast, closed-form, and efficient estimators for hierarchical models with AR(1) covariance and unequal cluster sizes, Communications in Statistics - Simulation and Computation, DOI: [10.1080/03610918.2017.1316395](https://doi.org/10.1080/03610918.2017.1316395)

To link to this article: <http://dx.doi.org/10.1080/03610918.2017.1316395>

Accepted author version posted online: 20 Apr 2017.
Published online: 20 Apr 2017.

Submit your article to this journal [↗](#)

Article views: 20

View related articles [↗](#)

View Crossmark data [↗](#)

Fast, closed-form, and efficient estimators for hierarchical models with AR(1) covariance and unequal cluster sizes

Lisa Hermans^a, Vahid Nassiri^b, Geert Molenberghs^{a,b}, Michael G. Kenward^c,
Wim Van der Elst^{a,d}, Marc Aerts^a, and Geert Verbeke^{a,b}

^aI-BioStat, Universiteit Hasselt, Diepenbeek, Belgium; ^bI-BioStat, KU Leuven, Leuven, Belgium;

^cLuton, United Kingdom; ^dJanssen Pharmaceutica, Beerse, Belgium

ABSTRACT

This article is concerned with statistically and computationally efficient estimation in a hierarchical data setting with unequal cluster sizes and an AR(1) covariance structure. Maximum likelihood estimation for AR(1) requires numerical iteration when cluster sizes are unequal. A near optimal non-iterative procedure is proposed. Pseudo-likelihood and split-sample methods are used, resulting in computing weights to combine cluster size specific parameter estimates. Results show that the method is statistically nearly as efficient as maximum likelihood, but shows great savings in computation time.

ARTICLE HISTORY

Received 10 May 2016

Accepted 2 April 2017

KEYWORDS

Maximum likelihood;
Pseudo-likelihood; Unequal
cluster size

MATHEMATICS SUBJECT CLASSIFICATION

62B05; 62F10

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) provide 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 article 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 et al. (2003); Benhin et al. (2005); Hoffman et al. (2001); Cong et al. (2007); Chiang and Lee (2008); Wang et al. (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 modeling of outcomes and cluster size is not required.

Hermans et al. (2017) 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 covari-

ance d . Hermans et al. (2017) 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., 2017). 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. (2017) 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 article, 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 to 128 per country with a total of 2,039. The positive and negative syndrome scale (PANSS) was used to assess 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 article is organized 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 analyzed in Section 7. The closing

discussion is presented in [Section 8](#). Some additional background material is available in a separate, web-based appendix.¹

2. Model formulation

Suppose that there is a sample of N independent clusters, among which K different cluster sizes n_k ($k = 1, \dots, K$) 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 i th ($i = 1, \dots, c_k$) replicate among the clusters of size n_k by $\mathbf{Y}_i^{(k)}$.

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

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

$$\mathbf{b}_i^{(k)} \sim N(0, D), \quad (2)$$

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 $Z_i^{(k)} \mathbf{b}_i^{(k)}$ 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 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.1.

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

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

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

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

The score equation for the mean produces, as usual:

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

¹ Available at <https://ibiostat.be/online-resources>.

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, β is estimated by the OLS estimator. However, in our AR(1) case, solving the score equations leads to

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

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

$$\tilde{\mu} = \frac{1}{cn} \sum_{i=1}^c \sum_{j=1}^n Y_{ij}. \quad (6)$$

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

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

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

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

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

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

with

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

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

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

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

$$f(\rho) = (n-1)S_2\rho^3 - (n-2)R\rho^2 - (nS_2 + S_1)\rho + nR = 0, \quad (10)$$

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

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

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

$$\hat{\sigma}^2 = \frac{1}{c} \cdot \frac{1}{1 - \hat{\rho}^2} (S_1 + \hat{\rho}^2 S_2 - 2\hat{\rho}R), \quad (12)$$

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

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

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, $\hat{\rho} = t_1 - b/(3a)$. 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 A.2, lead to

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

It is convenient to slightly change (13) to

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

yielding a very simple inverse:

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

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, which 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. (2017) 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 two 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 $\mathbf{k}$ is complete for a parameter $\boldsymbol{\theta}$ in an 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:

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

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 (1982), Papadakis proposed, as early as 1937, a correction to the least-square estimator for correlated observations arising in such settings as adjoining plots designs (Bartlett, 1938, 1976, 1978; Papadakis, 1937). 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 (1982) 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 (1982). 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 4.1. *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 A.2.4. This result 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 4.2. *The result of Proposition 4.1 easily generalizes to a mean of the form $\mu = X\beta$, when the design is constant among clusters.*

5. Clusters of variable size

Hermans et al. (2017) 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

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

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

Hermans et al. (2017) showed that the sum of the weight matrices should be the identity matrix, an obvious result, and considered, among others, the optimal expression:

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

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

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

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

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

The optimal weight is then

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

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 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, that is, $c_k n_k$.

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

$$\tilde{\mu} = \sum_{k=1}^K a_k \mu_k, \quad (23)$$

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

$$\text{Var}(\tilde{\mu}) = \frac{\sum_{k=1}^K a'_k \sigma_k^2 (1 + \rho_k)}{\left(\sum_{k=1}^K a'_k\right)^2} + \frac{\sum_{k=1}^K \left[c_k(n_k - 2) \sum_{m=1}^K a'_m (\mu_k - \mu_m) \right]^2 \frac{1 - \rho_k^2}{c_k(n_k - 1)}}{\left(\sum_{k=1}^K a'_k\right)^4}. \quad (24)$$

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

$$\text{Var}(\tilde{\mu}) = \sigma^2(1 + \rho) \left\{ \frac{1}{\sum_{k=1}^K a'_k} \right\} + (1 - \rho^2) \left\{ \frac{\sum_{k=1}^K \frac{c_k(n_k - 2)^2}{(n_k - 1)} \left[\sum_{m=1}^K a'_m (\mu_k - \mu_m) \right]^2}{\left(\sum_{k=1}^K a'_k\right)^4} \right\}. \quad (25)$$

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 datasets, as follows:

1. Compute $\hat{\mu}_k$, $\hat{\sigma}_k^2$, and $\hat{\rho}$, 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 Appendices A.2.5 and A.2.6.

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. (2017).

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 $\hat{\theta}_k$, 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 that are obtained by substituting the unknown parameter by its estimate using size-proportional weights.

It is observed that, for $\hat{\mu}$ and $\hat{\sigma}^2$, using $\tilde{\theta}$ in optimal weights does not increase the variance to a noticeable degree, but the effect for $\hat{\rho}$ 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 A.3.

7. Application: Clinical trials in schizophrenia

The data, introduced in Section 1, are analyzed here. The active treatments are: risperidone, haloperidol, perphenazine, and zuclopernthixol. We included for analysis patients with at least one follow-up measurement. Table 1 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:

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

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.

n	Pattern	FIN-1	FRA-3	Trial INT-2	INT-3	INT-7	Total
2	(0, 1)	17	8	71	43	3	142
	(0, 2)	0	0	2	0	1	3
	(0, 4)	0	0	1	0	0	1
3	(0, 1, 2)	8	4	83	41	7	143
	(0, 2, 4)	0	0	2	0	0	2
	(0, 1, 4)	1	0	3	1	0	5
4	(0, 1, 2, 4)	11	0	85	66	5	167
	(0, 2, 4, 6)	0	0	1	0	1	2
	(0, 2, 4, 8)	0	0	1	0	0	1
5	(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
6	(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
7	(0, 1, 4, 6, 8)	0	0	6	0	0	6
	(0, 1, 2, 6, 8)	0	0	8	0	0	8
	(0, 2, 4, 6, 8)	0	0	3	0	2	5
8	(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
9	(0, 1, 2, 4, 6, 8)	0	0	986	240	74	1,300
	(0, 1, 4, 6, 8, 10)	0	0	1	0	0	1
	(0, 1, 2, 6, 8, 12)	0	0	1	0	0	1
10	(0, 1, 2, 4, 6, 10)	0	0	1	0	0	1
	(0, 1, 2, 4, 5, 6)	0	0	2	0	0	2

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 that 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 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	✓	✓	✓	✓	✓

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)

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 estimates; their confidence intervals largely overlap.

As mentioned previously, these data are assembled from five 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:

$$Y_{ijk} = \mu + \tau_i + \alpha_j + \beta t_{ij} + (\tau\alpha)_{ij} + (\tau\beta)_{ik} + (\alpha\beta)_{jk} + (\tau\alpha\beta)_{ijk} + \epsilon_{ijk},$$

$$i = 1, \dots, 5, \quad j = 1, \dots, 4, \quad k = 1, \dots, n, \quad \epsilon_{ijk} \sim N_n(0, R), \quad (27)$$

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 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 A.4.

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)

8. Concluding remarks

As an extension to the normal-compound symmetry model, discussed in Hermans et al. (2017), 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, which 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 parameter-free and simple to use. Simulations show that in certain large settings, computation time can be 1,000 times faster than with standard maximum likelihood.

There are missing observations in the PANSS dataset. 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 article 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 No. 602552. The authors gratefully acknowledge support from the IWT-SBO ExaScience grant.

References

- Aerts, M., Faes, C., Hens, N., Loquiha, O., Molenberghs, G. (2011). Incomplete clustered data and non-ignorable cluster size. In: Conesa, D., Forte, A., López-Quílez, A., Muñoz, F., eds. *Proceedings of the 26th International Workshop on Statistical Modelling*. València, Spain, pp. 35–40.
- Aerts, M., Geys, H., Molenberghs, G., Ryan, L. (2002). *Topics in Modelling of Clustered Data*. London: Chapman & Hall.
- Alonso, A., Geys, H., Molenberghs, G., 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., Scott, A. J. (2005). Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes. *Biometrika* 92:435–450.
- Casella, G., Berger, R. L. (2001). *Statistical Inference*. Pacific Grove: Duxbury Press.
- Chiang, C-T., Lee, K-Y. (2008). Efficient estimation methods for informative cluster size data. *Statistica Sinica* 18:121–133.
- Cochran, W. G., Bliss, C. I. (1948). Discriminant function with covariance. *Annals of Mathematical Statistics* 19:151–176.
- Cong, X-J., Yin, G., Shen, Y. (2007). Marginal analysis of correlated failure time data with informative cluster sizes. *Biometrics* 63:663–672.
- Fieuws, S., 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., Verbeke, G. (2017). A tutorial on the practical use and implication of complete sufficient statistics.
- Hermans, L., Nassiri, V., Molenberghs, G., Kenward, M. G., Van der Elst, W., Aerts, M., Verbeke, G. (2017). Clusters with unequal size: Maximum likelihood versus weighted estimation in large samples. *Statistica Sinica*.
- Hoffman, E. B., Sen, P. K., 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*. Unpublished doctoral dissertation, UK.
- Koopmans, T. (1942). Serial correlation and quadratic forms in normal variables. *Annals of Mathematical Statistics* 13:14–33.
- Liu, A., Hall, W. J. (1999). Unbiased estimation following a group sequential test. *Biometrika* 86:71–78.
- Liu, A., Hall, W. J., Yu, K. F., Wu, C. (2006). Estimation following a group sequential test for distributions in the one-parameter exponential family. *Statistica Sinica* 16:165–181.
- Martin, R. J. (1982). Papadakis method. In: *Encyclopedia of Statistical Science*, Samuel, K., Normal L. J. (eds), USA: John Wiley and Sons Inc., Vol. 6., pp. 565:568.
- Milanzi, E., Molenberghs, G., Alonso, A., Kenward, M. G., Tsiatis, A., Davidian, M., 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., 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., 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 Salinique* 23:1–30.
- Shelbey, S. (1975). *CRC Standard Mathematical Tables*. Boca Raton, FL: CRC Press.
- Wang, M., Kong, M., 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., Satten, G. A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics* 59:36–42.