Statistica Sinica **28** (2018), 1107-1132
doi:https://doi.org/10.5705/ss.202016.0019

CLUSTERS WITH UNEQUAL SIZE: MAXIMUM LIKELIHOOD VERSUS WEIGHTED ESTIMATION IN LARGE SAMPLES

Lisa Hermans[1], Vahid Nassiri[2], Geert Molenberghs[1,2], Michael G. Kenward[3], Wim Van der Elst[4], Marc Aerts[1] and Geert Verbeke[1,2]

[1] *Universiteit Hasselt,* [2] *KU Leuven,* [3] *Former London School of Hygiene and Tropical Medicine and* [4] *Janssen Pharmaceutica*

Abstract: The analysis of hierarchical data that take the form of clusters with random size has received considerable attention. The focus here is on samples that are very large in terms of number of clusters and/or members per cluster, on the one hand, as well as on very small samples (e.g., when studying rare diseases), on the other. Whereas maximum likelihood inference is straightforward in medium to large samples, in samples of sizes considered here it may be prohibitive. We propose sample-splitting (Molenberghs, Verbeke and Iddi (2011)) as a way to replace iterative optimization of a likelihood that does not admit an analytical solution, with closed-form calculations. We use pseudo-likelihood (Molenberghs et al. (2014)), consisting of computing weighted averages over solutions obtained for each cluster size occurring. As a result, the statistical properties of this approach need to be investigated, especially because the minimal sufficient statistics involved are incomplete. The operational characteristics were studied using simulations. Simulations were also done to compare the proposed method to existing techniques developed to circumvent difficulties with unequal cluster sizes, such as multiple imputation. It follows that the proposed non-iterative methods have a strong beneficial impact on computation time; at the same time, the method is the most precise among its competitors considered. The findings are illustrated using data from a developmental toxicity study, where clusters are formed of fetuses within litters.

Key words and phrases: Likelihood inference, pseudo-likelihood, unequal cluster size.

1. Introduction

Much statistical theory is derived under the paradigm of a fixed sample size. However, there are many common practical settings in which this paradigm does not hold. Examples include sequential trials, where the trial may be stopped

early at a number of time points during accrual, because of the strength, or lack, of a treatment effect; incomplete data in longitudinal studies or surveys; longitudinal data with random measurement occasions; and censored survival data. Molenberghs et al. (2014) provide an overview of such situations.

Here, we focus on hierarchical (or clustered) data with unequal cluster sizes.

Clustering is taken in its broadest sense, encompassing longitudinal data, family-based studies, toxicology (Aerts et al. (2002)), agricultural experiments, multi-level designs in the social and behavioral sciences, and so on. In longitudinal trials, it is not uncommon to plan for the same number of measurements to be taken per study subject, often at a common set of time points. If all data were collected according to protocol, the cluster size would be fixed. However, even in such studies, cluster sizes are often *de facto* random because of incompleteness in the data. In many random cluster size settings there may be associations between outcomes and cluster size. In part of the literature, this is termed 'informative cluster size' and a suite of methods has been proposed to accommodate this situation, many based on inverse probability weighting (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)). Unequal cluster sizes can occur for any outcome type, including continuous, binary, categorical, count, and event time.

Unequal cluster sizes may or may not be governed by a stochastic mechanism. For example, they can be unequal by design choice, without being stochastic; e.g., when a sample is selected in each town proportion to the population size. Litter sizes in pregnant rodents will truly be stochastic. When stochastic, the mechanism is completely random when it depends on neither observed nor unobserved data; it is random when it depends on observed but, given these, not on unobserved data; other mechanisms are termed non-random. In the literature, mechanisms other than complete random are often termed informative. Although an important issue, we do not focus on informative cluster sizes here. Attention is confined to the case where cluster size is unequal, but independent of both observed and unobserved outcomes. In doing so we distinguish issues that stem purely from the non-constant nature of the cluster size, from those that result from the association between cluster size and outcome. We focus on the differences between the case of a fixed cluster size that is common to all clusters, and that of a fluctuating cluster size, whether for design reasons or randomly. In particular, the joint modelling of outcomes and cluster size is not considered.

As a simple, yet non-trivial, clustering paradigm, we consider the normal

compound-symmetry (CS) model, which is a three-parameter multivariate normal model, with a common mean μ, a common variance $\sigma^2 + d$, and a common covariance d. Molenberghs, Verbeke and Iddi (2011) studied this case in the context of so-called split-sample methodology: they proposed a particular form of pseudo-likelihood where a sample is subdivided into M subsamples, which are separately analyzed as if they were unrelated, after which the results are averaged using appropriate weights, leading to proper point and precision estimates. Pseudo-likelihood has received considerable attention (Varin, Reid and Firth (2011); Molenberghs and Verbeke (2005, Chap. 9, 12, 21, 24, 25); Aerts et al. (2002, Chap. 6, 7)).

Assume that there are c_k clusters of size n_k, $k = 1, \ldots, K$. For ease of development, we allow for some of the n_k to be equal, which is useful when a subgroup of clusters that is of the same size is chosen to be sub-divided (because there are very many or for other reasons). A natural split is made with respect to the cluster size, i.e., as if every cluster size defines its own stratum.

Evidently, for medium to large sample sizes, full maximum likelihood or Bayesian inferences are statistically optimal and computationally feasible; hence, the work done here might be less relevant. However, with really big data, where the number of independent clusters runs in the millions or beyond, and/or in settings where the number of measurements per cluster becomes very large (e.g., in meta-analysis), maximum likelihood eventually becomes prohibitive in terms of computation time. At the other end of the spectrum, in very small samples (e.g., in small-area epidemiology applications, or when studies are conducted in so-called orphan diseases), maximum likelihood estimates may become unstable, to the point where it is difficult to obtain convergence. This may be due, for example, to relatively flat likelihood functions. The non-iterative nature of our proposal removes such issues. Small samples refers here to a small number of clusters; the clusters themselves may consist of smaller or larger numbers of within-cluster replication. We are not the first to consider these issues. Van der Elst et al. (2015) considered multiple imputation to bring clusters to the same size before applying maximum likelihood. If done with care, convergence problems are drastically reduced. Williamson, Datta and Satten (2003) and Follmann, Proschan and Leifer (2003) proposed so-called multiple outputation, to repeatedly create independent samples by randomly selecting one member per cluster. To ensure that correlation is taken into account, combination rules reminiscent of multiple imputation are then applied to combine inferences from the samples drawn. These methods are based on repeated sampling and come at

computational cost for high-dimensional data (Sikorska et al. (2013)). Therefore, in this paper, the focus is on entirely non-iterative methods, bringing together the advantages of balanced data and simple averaging methodology. A consequence of our approach is the need for applying weights when combining results from the K strata. We establish how results on incomplete sufficient statistics in the context of weighted averages (Molenberghs et al. (2014); Hermans et al. (2017)) imply that there may be no optimal set of weights. Given this, we propose pragmatically attractive weights, in terms of efficiency, bias, and computational ease.

The remainder of the paper is organized as follows. Two motivating datasets are described in Section 2. In Section 3 essential background material on incomplete sufficient statistics is presented. The compound-symmetry model is introduced in Section 4, and a review is provided of the relevant incompleteness results from Hermans et al. (2017), together with implications for likelihood-based estimation. Background from the pseudo-likelihood-based split-sample method is given in Supplementary Material Section S.4. A general split-sample approach to the CS model is provided in Section 5, and a number of specific but practically relevant cases are considered. Details about the specifics for the CS case are presented in Section 6. Section 7 is dedicated to a simulation study, examining situations for which there are no closed forms on the one hand, and studying numerical performance (speed and convergence) on the other. The data, described in Section 2, are analyzed in Section 8. Ramifications and recommendations for practice are offered in Section 9.

2. Developmental Toxicity Study Sets

Data from the Research Triangle Institute under contract to the National Toxicology Program of the U.S.A. (NTP data), are analyzed. These developmental toxicity studies investigate the effects in mice of three chemicals: di(2--ethylhexyl)phthalate (DEHP) (Tyl et al. (1988)) ethylene glycol (EG) (Price et al. (1985)), and diethylene glycol dimethyl ether (DYME) (Price et al. (1987)). The studies were conducted in timed-pregnant mice during the period of major organogenesis. The dams were sacrificed, just prior to normal delivery, and the status of uterine implantation sites recorded. The outcome of interest here is fetal weight. Summary data from the DEHP trial are presented in Table 1. The design for EG and DYME is similar. It is clear from the table that average litter size is depleted with increasing dose, as is the average weight.

Table 1. Developmental Toxicity Study (DEHP). Summary data by dose group.

	# dams with		# live	average	
dose	implants	viable implants	fetuses	litter size	weight
0 mg/kg/day	30	30	330	13.2	0.9483
44 mg/kg/day	26	26	288	11.1	0.9592
91 mg/kg/day	26	26	277	10.7	0.8977
191 mg/kg/day	24	17	137	8.1	0.8509
292 mg/kg/day	25	9	50	5.6	0.6906

3. Incomplete Sufficient Statistics

A statistic $k(Y)$ of a random variable Y, with Y belonging to a family P_θ, is complete if, for every measurable function $g(\cdot)$, independent of θ, $E[g\{k(Y)\}] = 0$ for all θ, implies that $P_\theta[g\{k(Y)\} = 0] = 1$ for all θ (Casella and Berger (2001)). The Lehman-Scheffé theorem (Casella and Berger (2001)) states that, if a statistic is unbiased, complete, and sufficient for a parameter θ, then it leads to the best mean-unbiased estimator for θ, while Basu's theorem (Basu (1955)) has it that statistic that is both boundedly complete and sufficient is independent of any ancillary statistic. As has been shown in the sequential trial context, a lack of completeness does not preclude the existence of estimators with very good 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 entire exponential family. Molenberghs et al. (2014) and Milanzi et al. (2016) broadened it further to a stochastic rather than a deterministic stopping rule, hence encompassing the case of a completely random sample size. Indeed, it would seem at first sight that this latter case is standard, because the sample size is unrelated to the data, whether observed or not. Yet, even in this case, completeness no longer holds. What is more, incompleteness holds when the cluster size is non-constant for whatever reason.

4. The Compound-Symmetry Model

Let $\boldsymbol{Y}$ be a vector of length n, with $\boldsymbol{Y} \sim N(\mu \mathbf{1}_n, \sigma^2 I_n + dJ_n)$. In general, both $\boldsymbol{Y}$ and n are random variables.

Suppose that there is a sample of N independent clusters, among which K different cluster sizes n_k $(k = 1, \ldots, K)$ are distinguished. Let the multiplicity of cluster size n_k be equal to c_k. Evidently, $N = \sum_{k=1}^{K} c_k$. Denote the outcome

vector for the ith ($i = 1, \ldots, c_k$) replicate among the clusters of size n_k by $\boldsymbol{Y}_i^{(k)}$. We first show incompleteness of the sufficient statistic, then turn to likelihood estimation. For both, we start from the log-likelihood function.

4.1. Incompleteness

The data-dependent terms in the log-likelihood can be written as:

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{i=1}^{c_k} -\frac{1}{2}\left(\boldsymbol{Y}_i^{(k)} - \mu\mathbf{1}_{n_k}\right)' \left(\sigma^2 I_{n_k} + dJ_{n_k}\right)^{-1} \left(\boldsymbol{Y}_i^{(k)} - \mu\mathbf{1}_{n_k}\right) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{c_k} -\frac{1}{2}\left(\boldsymbol{Y}_i^{(k)} - \mu\mathbf{1}_{n_k}\right)' \left(I_{n_k} - \frac{d}{\sigma^2 + n_k d}J_{n_k}\right) \left(\boldsymbol{Y}_i^{(k)} - \mu\mathbf{1}_{n_k}\right) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{c_k} \frac{\mu}{\sigma^2 + n_k d}\left(\sum_{j=1}^{n_k} Y_{ij}^{(k)}\right) - \frac{1}{2\sigma^2}\left(\sum_{k=1}^{K}\sum_{i=1}^{c_k}\sum_{j=1}^{n_k} Y_{ij}^{(k)2}\right) \\
&\quad + \sum_{k=1}^{K}\sum_{i=1}^{c_k} \frac{d}{2\sigma^2(\sigma^2 + n_k d)}\left(\sum_{j=1}^{n_k} Y_{ij}^{(k)}\right)^2. \qquad (4.1)
\end{aligned}$$

The three terms in (4.1) are qualitatively different. Indeed, the middle one corresponds to a single sufficient statistic, the sum of all squares across clusters, while the first and last split into as many sufficient statistics as there are unique cluster sizes.

Hermans et al. (2017) proved a characterization of incompleteness, essentially stating that when the dimension of the sufficient statistic is larger than the dimension of the parameter vector, the sufficient statistic is no longer complete. More details can be found in Supplementary Materials Section S1. This sharp division also occurs when studying certain properties of the maximum likelihood estimator.

4.2. Likelihood-based estimation of the CS model

Similar in spirit to (4.1), but now using all terms, the log-likelihood can be written as

$$\ell(\mu, \sigma^2, d) = \sum_{k=1}^{K} \ell_k(\mu, \sigma^2, d), \qquad (4.2)$$

with the cluster size specific log-likelihood term

$$\ell_k(\mu, \sigma^2, d) = -\frac{1}{2}\sum_{i=1}^{c_k}\left[\ln\left\{\sigma^{2n_k} + n_k\sigma^{2(n_k-1)}d\right\}\right.$$

$$+ \left(\boldsymbol{Y}_i^{(k)} - \mu \mathbf{1}_{n_k}\right)' \frac{1}{\sigma^2} \left(I_{n_k} - \frac{d}{\sigma^2 + n_k d} J_{n_k} \right) \left(\boldsymbol{Y}_i^{(k)} - \mu \mathbf{1}_{n_k}\right) \Bigg]. \tag{4.3}$$

Using derivations similar to those in Molenberghs, Verbeke and Iddi (2011), the cluster size specific log-likelihood can be maximized analytically *assuming that there is a separate parameter per cluster size*. By replacing $\ell_k(\mu, \sigma^2, d)$ by $\ell_k(\mu_k, \sigma_k^2, d_k)$, we can consider the kernel of the log-likelihood, in general for K cluster sizes, and allowing for the parameter vector to change with cluster size:

$$\ell\left(\{\mu_k\}_k, \{\sigma_k^2\}_k, \{d_k\}_k\right) \propto -\frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{c_k} \Big\{ \ln |\Sigma_{n_k}| + \left(\boldsymbol{y}_i^{(k)} - \boldsymbol{\mu}_{n_k}\right)' \Sigma_{n_k}^{-1} \left(\boldsymbol{y}_i^{(k)} - \boldsymbol{\mu}_{n_k}\right) \Big\}, \tag{4.4}$$

where $\boldsymbol{\mu}_k = \mu_k \mathbf{1}_{\boldsymbol{n_k}}$ $\Sigma_{n_k} = \sigma_k^2 I_{n_k} + d_k J_{n_k}$. The score functions are presented in Supplementary Materials Section S.2. Solving these score functions (S2.1)-(S2.3) leads to:

$$\hat{\mu}_k = \frac{1}{c_k n_k} \sum_{i=1}^{c_k} \sum_{j=1}^{n_k} Y_{ij}^{(k)}, \tag{4.5}$$

$$\hat{\sigma}_k^2 = \frac{1}{c_k n_k (n_k - 1)} \left(n_k \sum_{i=1}^{c_k} \boldsymbol{Z}_i^{(k)'} \boldsymbol{Z}_i^{(k)} - \sum_{i=1}^{c_k} \boldsymbol{Z}_i^{(k)'} J_{n_k} \boldsymbol{Z}_i^{(k)} \right), \tag{4.6}$$

$$\hat{d}_k = \frac{1}{c_k n_k (n_k - 1)} \left(\sum_{i=1}^{c_k} \boldsymbol{Z}_i^{(k)'} J_{n_k} \boldsymbol{Z}_i^{(k)} - \sum_{i=1}^{c_k} \boldsymbol{Z}_i^{(k)'} \boldsymbol{Z}_i^{(k)} \right), \tag{4.7}$$

where $\boldsymbol{Z}_i^{(k)} = (\boldsymbol{Y}_i^{(k)} - \mu_k \mathbf{1}_{n_k})$.

When the cluster size is constant, the compound-symmetry model has closed form ML estimators, given by (4.5)–(4.7). Closed-form estimators for the variance-covariance matrix of the estimator exist as well (Molenberghs, Verbeke and Iddi (2011)). For the mean, the variance is:

$$\text{var}(\hat{\mu}_k) = \frac{\sigma_k^2 + n_k d_k}{c_k n_k}. \tag{4.8}$$

The expressions for the variance-covariance structure of $(\hat{\sigma}_k^2, \hat{d}_k)$ is:

$$\text{var} \begin{pmatrix} \hat{\sigma}_k^2 \\ \hat{d}_k \end{pmatrix} = \frac{2\sigma_k^4}{c_k n_k (n_k - 1)} \begin{pmatrix} n_k & -1 \\ -1 & \dfrac{\sigma_k^4 + 2(n_k - 1) d_k \sigma_k^2 + n_k (n_k - 1) d_k^2}{\sigma_k^4} \end{pmatrix}. \tag{4.9}$$

The mean parameter is independent of the variance components.

These results can be used when a separate parameter vector is estimated for each of the cluster sizes and, as a special case, when there is only one cluster

size. Four features of use in what follows are: (a) there are closed forms; (b) the sufficient statistic is complete; (c) the estimator is unique minimum variance unbiased; (d) the mean parameter estimator and the variance parameter estimator are independent.

These results are lost when $K \geq 2$. We briefly sketch the lack of closed-form solutions in this case in Supplementary Materials Section S2.2.

The lack of a closed form is well known, but we highlight a few relevant features here. More detail is given in Supplementary Materials Section S3, where we show

$$\hat{\mu} = \frac{\sum_{k=1}^{K}\{(n_k c_k)/(\sigma^2 + n_k d)\}\hat{\mu}_k}{\sum_{k=1}^{K}\{(n_k c_k)/(\sigma^2 + n_k d)\}}. \tag{4.10}$$

Examining (4.10) suggests weighted averages:

$$\tilde{\mu} = \sum_{k=1}^{K} a_k \hat{\mu}_k, \qquad \tilde{\sigma}^2 = \sum_{k=1}^{K} b_k \hat{\sigma}_k^2, \qquad \tilde{d} = \sum_{k=1}^{K} g_k \hat{d}_k. \tag{4.11}$$

This idea is similar to that in Molenberghs, Verbeke and Iddi (2011), who split a sample in sub-samples, analyzed each separately, and then combined the result in an overall estimator. They considered splits in both dependent and independent sub-samples. Dependent samples occur when very long sequences of repeated measures are collected, which are then sub-divided for convenience. This approach is not of use here. Independent samples arise when there are many independent replicates, i.e., a large number of clusters. They studied the CS case, but only for a single cluster size. The total number of clusters was then split into M parts comprising an equal number of clusters. We modify these ideas to the case of unequal cluster sizes, with a variable number of clusters per split.

5. Split-sample Methods for Clusters of Variable Size

The derivations are based on general pseudo-likelihood principles, reviewed in Supplementary Materials Section S4. We first make generic the setting at the beginning of Section . Let there be a sample of N independent clusters. Partition the sample into K sub-samples, with c_k independent and identically distributed clusters in sub-sample k; $N = \sum_{k=1}^{K} c_k$. $\boldsymbol{Y}_i^{(k)}$ remains the outcome vector for replicate i in sub-sample k.

Subjects in different sub-samples are allowed to have the same distribution, but subjects in the same sub-sample *must* have the same distribution. This covers the running example of CS clusters, partitioned according to cluster size. However, it is possible to further sub-divide such a sub-sample in various sub-

samples, all with the same cluster size. This is sensible, for example, in very large databases. An extreme example follows when sub-samples consist of a single independent replicate, useful, for example, in a meta-analysis with large individual studies. This limiting situation can also be considered with CS data, because all clusters (except those of size 1) contribute to all three parameters.

Consider pseudo-likelihood in this general case (see also Eq. (S4.4)). Assume that $\boldsymbol{\theta}^*$ is a vector of length p, and that each $\boldsymbol{\theta}_k$ is a separate copy of $\boldsymbol{\theta}^*$. Then it can be shown that the generic combination rules are

$$\tilde{\boldsymbol{\theta}}^* = \sum_{k=1}^{K} A_k \hat{\boldsymbol{\theta}}_k, \tag{5.1}$$

$$\mathrm{var}(\tilde{\boldsymbol{\theta}}^*) = \sum_{k=1}^{K} A_k V_k A_k', \tag{5.2}$$

with $V_k = I_0(\boldsymbol{\theta}_k)^{-1}$. We use the symbol $\tilde{\boldsymbol{\theta}}^*$ to emphasize that this is not necessarily the maximum likelihood estimator even though, in our formalism, $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimator when restricting attention to sub-sample k. Equation (5.2) is appropriate only when the weights A_k are free of the parameters to be estimated. We return to this at the end of the section.

Weighting Schemes Not every choice of the A_k leads to an unbiased estimator. To enforce unbiasedness, consider the requirement

$$\boldsymbol{\theta} = E(\hat{\boldsymbol{\theta}}^*) = \sum_{k=1}^{K} A_k E(\hat{\boldsymbol{\theta}}_k) = \left(\sum_{k=1}^{K} A_k \right) \boldsymbol{\theta},$$

whence $I_p = \sum_{k=1}^{K} A_k$. This requirement is satisfied for (S4.5). This suggests two obvious choices:

Constant weights. Set $A_k = (1/K) I_p$.

Proportional weights. Set $A_k = (c_k/N) I_p$.

Constant weights are the clear choice when all subjects are i.i.d. and partitioning is in sub-samples of equal size. Proportional weights are called for in the i.i.d. case, with sub-samples of varying size.

Consider optimal weights through the objective function

$$Q = \sum_{k=1}^{K} A_k V_k A_k' - \Lambda \left(\sum_{k=1}^{K} A_k - I_p \right),$$

where Λ is a matrix of Lagrange multipliers. Taking the first derivative of Q w.r.t. A_k leads to $A_k = \Lambda V_k^{-1}/2$. Because the A_k sum to the identity, $\Lambda =$

$2(\sum_{m=1}^{K} V_m^{-1})^{-1}$ and we have the following.

Optimal weights. These take the form:

$$A_k^{\text{opt}} = \left(\sum_{m=1}^{K} V_m^{-1} \right)^{-1} V_k^{-1}. \tag{5.3}$$

With this choice, (5.1)–(5.2) become:

$$\tilde{\boldsymbol{\theta}}^* = \hat{\boldsymbol{\theta}}^* = \left(\sum_{k=1}^{K} V_k^{-1} \right)^{-1} \sum_{k=1}^{K} V_k^{-1} \hat{\boldsymbol{\theta}}_k, \tag{5.4}$$

$$\text{var}(\tilde{\boldsymbol{\theta}}^*) = V = \left(\sum_{k=1}^{K} V_k^{-1} \right)^{-1}. \tag{5.5}$$

The optimal weights lead to the maximum likelihood estimator. To apply the optimal weights in practice is typically not straightforward. A closed form expression for the V_k does not always exist, and even if it did, as in the CS case, it may depend on the unknown parameters. The optimal weights can suggest sensible choices and we describe a couple of these. They will be illustrated in the next section for the CS case.

Scalar weights. While optimal weights may be unwieldy, one could consider scalar weights by requiring the A_k to be diagonal. This implies that each component of $\boldsymbol{\theta}^*$, θ_r^*, say, is a linear combination

$$\tilde{\theta}_r^* = \sum_{k=1}^{K} a_{k,r} \hat{\theta}_{k,r},$$

where then, formally, $A_k = \text{diag}(a_{k,1}, \ldots, a_{k,p})$. The optimization route, followed for unrestricted A_k, can then be followed component-wise as well. Because the class of A_k over which to optimize is restricted, the resulting optimum does not necessarily correspond to the maximum likelihood solution. The rationale for choosing this route is computational convenience, and its advantages vary from problem to problem.

Iterated optimal weights. An iterative scheme can be followed:

1. Estimate $\hat{\boldsymbol{\theta}}_k$.
2. Compute an initial estimator for $\boldsymbol{\theta}^*$, $\boldsymbol{\theta}^{*(0)}$, say, using a simple weighting method, e.g., using constant or proportional weights.
3. Using the current parameter estimate, $\boldsymbol{\theta}^{*(t)}$ say, calculate $V_k^{(t+1)}$.
4. Determine:

$$\boldsymbol{\theta}^{*(t+1)} = \left[\sum_{k=1}^{K}\left\{V_k^{(t+1)}\right\}^{-1}\right]^{-1}\sum_{k=1}^{K}\left\{V_k^{(t+1)}\right\}^{-1}\hat{\boldsymbol{\theta}}_k.$$

5. Repeat steps 2–3 until convergence.

This scheme can always be followed and it has the advantage that the data need only be analyzed once, to yield $\hat{\boldsymbol{\theta}}_k$. From this point on, calculations involve algebraic expressions for the parameters only.

Approximate optimal weighting. Related to the previous method, a non-iterative approximation consists of replacing V_k by $V_k(\tilde{\boldsymbol{\theta}}_k)$ in (5.4). Here, $\tilde{\boldsymbol{\theta}}_k$ could be, for example, the sub-sample specific estimator $\hat{\boldsymbol{\theta}}_k$, or the $\tilde{\boldsymbol{\theta}}^*$ obtained using a simple scheme, such as constant or proportional weighting. This method avoids all further iteration, once the $\boldsymbol{\theta}_k$ have been determined.

Approximate optimal weighting is a method that could be considered when the use of (5.2) might lead to underestimation of the variability, because the A_k now depend on the parameters estimated from stratum k. To properly account for this extra source of uncertainty. Consider that

$$\frac{\partial}{\partial\boldsymbol{\theta}_k}(A_k\boldsymbol{\theta}_k) = A_k + \left(\frac{\partial A_k}{\partial\theta_{k1}}\boldsymbol{\theta}_k\middle|\cdots\middle|\frac{\partial A_k}{\partial\theta_{kp}}\,\boldsymbol{\theta}_k\right), \tag{5.6}$$

where θ_{kj}, $j = 1,\ldots,p$ ranges over the components of $\boldsymbol{\theta}_k$. Writing $W_k = V_k^{-1}$ for ease of notation,

$$\frac{\partial A_k}{\partial\theta_{kj}} = W^{-1}\frac{\partial W_k}{\partial\theta_{kj}}(I_p - W^{-1}W_k), \tag{5.7}$$

with I_p the p-dimensional identity matrix. Plugging (5.7) into (5.6), the proper delta-method approximation to the variance is

$$\mathrm{var}(\tilde{\boldsymbol{\theta}}^*) \simeq \sum_{k=1}^{K}(A_k + B_k)V_k(A_k + B_k)', \tag{5.8}$$

with

$$B_k = (\mathbf{1}_p'\otimes I_p)(I_p\otimes W^{-1})\mathrm{diag}\left(\frac{\partial W_k}{\partial\theta_{k1}},\ldots,\frac{\partial A_k}{\partial\theta_{kp}}\right)\left\{I_p\otimes(I_p - W^{-1}W_k)\boldsymbol{\theta}\right\},$$

and $\otimes$ signifying Kronecker product.

6. Partitioned-Sample Analysis for the Compound-Symmetry Model

For the normal compound-symmetry model, a variety of options exists. We sketch them here, and then consider some in greater detail.

Consider the i.i.d. case, where all clusters are of the same size. Full maximum likelihood then leads to a closed-form solution. Molenberghs, Verbeke and Iddi

(2011) studied splitting the sample in dependent sub-samples for this case, and showed that splitting leads to efficiency loss for the variance components, but not for the mean. They split the sequences of repeated measures in portions of equal size. Unequally sized splits could also be considered, although the rationale for this may not be compelling. They did not consider splits in independent sub-samples. We do so here, in Section 6.2, both for sub-samples of equal as well as for unequal size.

With variable cluster size, we know from Section 4.2 that full maximum likelihood does not lead to a closed-form solution. We will study in more detail the natural splitting into sub-samples of constant cluster size.

A special case, for both the i.i.d. and unequal cluster-size settings, is the cluster-by-cluster analysis. We will apply our methodology, outlined in Section 5, to this case, and contrast it with an *ad hoc* moment-based set of estimators.

6.1. Variable cluster size

6.1.1. Optimal weights

As we see in Section 6.1.3, scalar and optimal (hence, vectorized) weights do not make a difference for the mean parameter, because of the independence between the mean and the covariance parameters.

We can therefore consider the mean parameter separately from the covariance parameters. Let v_k be the variance of the mean in stratum k, and V_k the corresponding variance-covariance matrix for the variance components. Applying optimal weight (5.3) to the mean produces

$$\tilde{\mu} = \left(\sum_{k=1}^{K} \frac{c_k n_k}{\hat{\sigma}_k^2 + n_k \hat{d}_k} \right)^{-1} \sum_{k=1}^{K} \frac{c_k n_k \hat{\mu}_k}{\hat{\sigma}_k^2 + n_k \hat{d}_k}. \tag{6.1}$$

The corresponding estimators for the variance components, specific to a cluster size, are given by (4.6) and (4.7). Using them, and expression (4.9) for the variance, it follows that the optimal weighted estimator satisfies

$$\begin{pmatrix} \tilde{\sigma}^2 \\ \tilde{d} \end{pmatrix} = \left(\sum_{k=1}^{K} V_k^{-1} \right)^{-1} \sum_{k=1}^{K} \left\{ \begin{array}{c} \dfrac{Q_k}{2\hat{\sigma}_k^2} - \dfrac{d_k(2\hat{\sigma}_k^2 + n_k \hat{d}_k)}{2\hat{\sigma}_k^4(\hat{\sigma}_k^2 + n_k \hat{d}_k)^2} R_k \\ \dfrac{R_k}{2(\sigma_k^2 + n_k d_k)^2} \end{array} \right\}, \tag{6.2}$$

with Q_k and R_k as in (S.9) and (S.10), respectively.

6.1.2. Iterated and approximate optimal weights

Evidently, the principles of iterated and approximate optimal weights can be applied here.

Replacing the variance components in (6.1) by their expectation leads to:

$$\tilde{\mu} = \left(\sum_{k=1}^{K} \frac{c_k n_k}{\sigma^2 + n_k d} \right)^{-1} \sum_{k=1}^{K} \frac{c_k n_k \hat{\mu}_k}{\sigma^2 + n_k d}. \tag{6.3}$$

If we do the same for the mean, on both sides of the equality, we obtain

$$\mu = \left(\sum_{k=1}^{K} \frac{c_k n_k}{\sigma^2 + n_k d} \right)^{-1} \sum_{k=1}^{K} \frac{c_k n_k \mu}{\sigma^2 + n_k d}. \tag{6.4}$$

Although (6.1) cannot directly be used, because of circularity, (6.3) and (6.4) are available to us.

Replacing the variance components on the right hand side of (6.2) by their expectations leads to

$$\begin{pmatrix} \tilde{\sigma}^2 \\ \tilde{d} \end{pmatrix} = \left(\sum_{k=1}^{K} V_k^{-1} \right)^{-1} \sum_{k=1}^{K} \begin{Bmatrix} \dfrac{Q_k}{2\sigma^2} - \dfrac{d(2\sigma^2 + n_k d)}{2\sigma^4(\sigma^2 + n_k d)^2} R_k \\ \dfrac{R_k}{2(\sigma^2 + n_k d)^2} \end{Bmatrix}. \tag{6.5}$$

Using their explicit expressions, and using the fact that the expectation must be $(\sigma^2, d)'$, (6.2) leads to the following identity:

$$\begin{pmatrix} \sigma^2 \\ d \end{pmatrix} = V \sum_{k=1}^{K} \frac{c_k n_k}{2(\sigma^2 + n_k d)} \begin{Bmatrix} \dfrac{\sigma^2 + (n_k - 1)d}{\sigma^2} \\ 1 \end{Bmatrix}. \tag{6.6}$$

Expressions (6.1) and (6.2) can be used for approximate weighting, by plugging in, as is done, on the right hand side, the cluster-size specific mean and variance components.

Expressions (6.3) and (6.5) can be used for iterated weighting. The estimator for the mean depends on the variance components, but not vice versa. This dependence is insightful: there is independence between mean and variance components for every cluster-size specific stratum separately. As a consequence, $\tilde{\mu}$ on the one hand, and $\tilde{\sigma}^2$ and $\tilde{d}$ on the other, can be determined separately, provided the latter are done first.

Expressions (6.4) and (6.6) move beyond the previous schemes and exist by virtue of their explicit expressions. In (6.6) an initial consistent estimator for the variance components can be used on the right hand side. Once the left hand

side has been determined, the result can be plugged in again on the right, until convergence. Once done, the final variance component estimates can be used in (6.4) and the process repeated for μ, until convergence.

6.1.3. Scalar weights

In this case, A_k equals $\text{diag}(a_k, b_k, g_k)$, with the scalars as in (4.11). Obviously, the conditions for unbiased estimators are $\sum_{k=1}^K a_k = \sum_{k=1}^K b_k = \sum_{k=1}^K g_k = 1$.

The stratum-specific estimators are given by (4.5)–(4.7) and their variance-covariance structure by (4.8)–(4.9). The objective function to find the optimum is

$$Q = \sum_{k=1}^K a_k^2 \text{var}(\hat{\mu}_k) - \lambda \left(\sum_{k=1}^K a_k \right).$$

Logic, similar to the vector case, and using the explicit expressions for the variances, leads to:

$$a_k = \frac{c_k n_k / (\sigma^2 + n_k d)}{\sum_{m=1}^K \{(c_m n_m)/(\sigma^2 + n_m d)\}} = \frac{c_k n_k / \{(1-\rho) + n_k \rho\}}{\sum_{m=1}^K \{(c_m n_m)/((1-\rho) + n_m \rho)\}}, \quad (6.7)$$

$$b_k = \frac{c_k (n_k - 1)}{\sum_{m=1}^K c_m (n_m - 1)}, \quad (6.8)$$

$$\begin{aligned} g_k &= \frac{c_k n_k / \{\sigma^4/(n_k - 1) + 2\sigma^2 d + n_k d^2\}}{\sum_{m=1}^K [c_m n_m / \{\sigma^4/(n_m - 1) + 2\sigma^2 d + n_m d^2\}]} \\ &= \frac{c_k n_k (n_k - 1)/[(1-\rho)^2 + \{2\rho(1-\rho) + n_k \rho^2\}(n_k - 1)]}{\sum_{m=1}^K (c_m n_m (n_m - 1)/[(1-\rho)^2 + \{2\rho(1-\rho) + n_m \rho^2\}(n_m - 1)])}, \end{aligned} \quad (6.9)$$

where $\rho = d/(\sigma^2 + d)$. Here, the coefficients depend on the parameters in different ways. While b_k is independent of the parameters, a_k has denominators linear in σ^2 and d (equivalently, in ρ), and g_k has quadratic functions instead.

These weights, like the optimal ones, depend on the parameters. Evidently, they can be made part of an iterative scheme, as with the vector-valued weights. The added advantages are that matrix computations simplify to scalar computations; for models with relatively few parameters, like the one here, this is a small advantage. More importantly, approximations can be considered for each parameter separately.

Direct calculations show that the variance for the weighted estimator of the mean, using weights (6.7), is equal to that of maximum likelihood. For this parameter, the weighted split-sample estimator is the maximum likelihood estimator, in spite of the use of the scalar weight. This is to be expected, because

V_k is block-diagonal and because of independence of the mean estimator from the variance components estimators within a given cluster size. This implies that the optimally weighted estimator and the scalar estimator coincide for the mean. They differ for the variance components.

6.1.4. Approximate optimal scalar weights

To illustrate the logic of this method, consider (6.7)–(6.9) for the case where cluster sizes, for a good majority of the clusters, are sufficiently large. Taking limits for $n_k \to +\infty$ produces

$$a_k^{\text{app}} = g_k^{\text{app}} = \frac{c_k}{N}. \tag{6.10}$$

When this approximation is sensible, the very simple proportional weights follow. These approximations are exact, for a_k and g_k, when $\rho = 1$. They deteriorate when ρ becomes smaller. For example, in case $\rho = 0$,

$$a_k(\rho = 0) = \frac{c_k n_k}{\sum_{m=1}^K c_m n_m},$$

$$g_k(\rho = 0) = \frac{c_k n_k (c_k n_k - 1)}{\sum_{m=1}^K c_m n_m (c_m n_m - 1)} \approx \frac{c_k^2 n_k^2}{\sum_{m=1}^K c_m^2 n_m^2}.$$

A reasonable approximation for b_k is

$$b_k^{\text{app}} = \frac{c_k n_k}{\sum_{m=1}^K c_m n_m}, \tag{6.11}$$

which sets it equal to $a_k(\rho = 0)$. The information for σ^2 is thus determined more in terms of the number of measurements than in the number of clusters. Dropping the n_k from this formula is sensible only when cluster sizes are not too different from one another.

Figure 1 depicts optimal scalar weights (6.7)–(6.9), alongside the apparently simplistic proportional weights, for two of the five NTP datasets chosen to represent two relatively different empirical cluster size distributions. In both cases, there is a considerable range of cluster sizes, approximately 1 to 20. At the same time, the frequencies of the cluster sizes vary considerably. The values for a_k and g_k are almost identical to the proportional weights. While a small discrepancy for b_k is noticeable, and understandable in view of (6.11), the proportional weights seem to offer a sensible choice. This issue will be examined further in the data-analytic Section 8.

6.2. The special case of common cluster size, splits of (un)equal size

When $n_k \equiv n$ is constant, (6.7)–(6.9) reduce to:

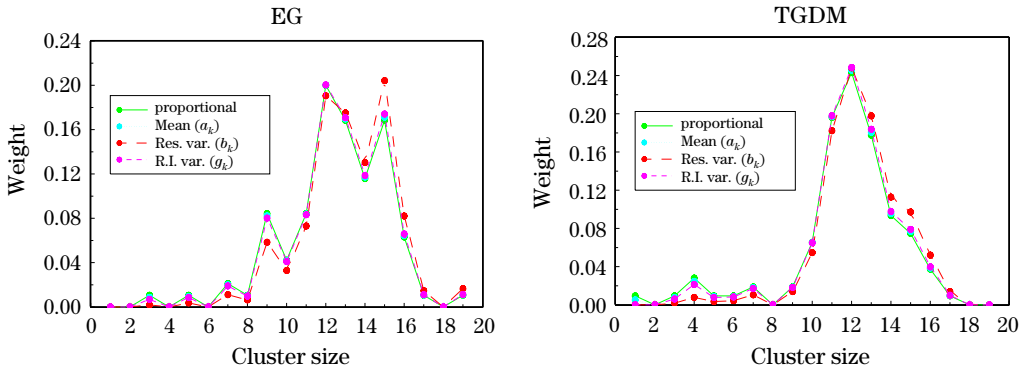

Figure 1. NTP Data. Scalar weights: proportional and optimal scalar versions for EG and TGDM datasets. The optimal scalar weights are computed for $\rho = d/(\sigma^2 + d) = 0.5$.

$$a_k = b_k = g_k = \frac{c_k}{N}, \tag{6.12}$$

Hence, while $a_k = b_k$ reduce to proportional weights, for g_k there is an impact of the partitioning structure. When, further, c_k is constant, we obtain $a_k = b_k = g_k = c/N = 1/K$, and equal weights follow. The similarities and the subtle differences with the results from Section 6.1.4 are worth pointing out. Expressions (6.10) and (6.12) are identical except for the parameter σ^2.

6.3. Cluster-by-cluster analysis

The expressions presented earlier in this section, using optimal weights and variations on this theme, can be applied when the partitioning is as extreme as possible: a single cluster per stratum. This sets $c_k \equiv 1$. Evidently, the n_k will then no longer be unique, but that is immaterial; while we make use of the fact that the cluster size is constant within a stratum, it does not need to be different between strata. We examine this case in more detail in Supplementary Materials Section S5.1. In particular, we derive under what asymptotics such an estimator is consistent.

7. Simulation Study

A first, limited, simulation study was carried out to examine the behavior of the partitioning method. Details are given in Supplementary Materials Section S6. Three settings were considered: (1) $c_k \cdot n_k$ is kept constant with the factors taking different values; (2) c_k is kept constant; (3) n_k is kept constant. For all of these, k goes from 1 to 4, so that there are four sub-samples. Apart from full likelihood, a series of weights was considered: equal, proportional to c_k,

size proportional to $c_k \cdot n_k$, approximate optimal, and iterated optimal.

From the results it is clear that equal weights are not a good choice. For μ and d, proportional weights are excellent, while for σ^2 so are the size proportional weights. Iterated optimal weights perform considerably better than approximate optimal weights, in the sense that the latter, like equal weights, arguably should not be considered for practice. When comparing iterated and approximate optimal weights, the former are more computationally intensive.

However, iterated optimal weights give results very close to proportional weights (for μ and d) and to size proportional weights (for σ^2). Importantly, all of these results are very close to the ones obtained from maximum likelihood.

As a consequence, we have a simple, non-iterative set of weights at our disposal, free of unknown parameters, with excellent performance.

A second simulation study compares the proposed methods to two alternatives: full maximum likelihood and multiple imputation. Details are reported in Supplementary Materials Section S7. The most striking conclusion is that closed-form solutions are much faster than their alternatives while, at the same time, yielding most precise results. The time gain of our fastest method relative to standard maximum likelihood using PROC MIXED ranges from 5 times to 30,000 times faster.

8. Analysis of Case Study

The data, introduced in Section 2, were analyzed in three ways. In Section 8.1 maximum likelihood estimators are presented, with split sampling, where splitting is by cluster size and using various weighting schemes. In Section 8.2, a dose effect is added to these. Finally, the cluster-by-cluster methodology of Section 6.3 is illustrated in Section 8.3.

8.1. Splitting by cluster size, no dose effect

Tables 2–4 present (restricted) maximum likelihood estimates (standard errors), together with those from various weighted estimators. The standard CS model is fitted to the fetal weight outcome, ignoring the dose effect. Because there is an effect of dose on litter size, the mean is associated with cluster size. It is therefore interesting to assess the impact of this on the split-sample estimators, when compared to the MLEs.

The ML and REML are very similar, with equal point estimates for μ, nearly equal estimates for σ^2, and similar estimates for d. The equality for the mean estimator is known for the CS case. The difference in the estimates of σ^2 arises

Table 2. NTP Data (DEHP). Cluster-by-cluster analysis. Maximum likelihood and weighted split-sample estimates (standard errors): (a) ML: maximum likelihood; (b) REML: restricted maximum likelihood; (c) Prop.: proportional weights; (d) Equal: equal weights; (e) Approx. sc.: like proportional weights, except that for b_k (6.11) is used; (f) Scalar: scalar weights, with the sub-sample specific weights plugged in for the parameters figuring in the weights; (g) Opt.: optimal weights, with the sub-sample specific weights plugged in for the parameters figuring in the weights. Proper: proper variances for optimal weights.

							Optimal	
Par.	ML	REML	Prop.	Equal	Approx. sc.	Scalar	Simpl.	Proper
Weighted [(S5.2)(S5.3)(S5.4)]								
μ	0.90718	0.90716	0.90602	0.89558	0.90602	0.92080	0.92080	
σ^2	0.01877	0.01877	0.02122	0.02244	0.01895	0.01871	0.01246	
d	0.01181	0.01195	0.00951	0.01016	0.00951	0.00085	0.00087	
s.e.($\hat{\mu}$)	0.01149	0.01155	0.01076	0.01360	0.01076	0.00766	0.00766	0.00766
s.e.($\hat{\sigma}^2$)	0.00084	0.00084	0.00128	0.00199	0.00094	0.00092	0.00061	0.00138
s.e.($\hat{d}$)	0.00196	0.00199	0.00210	0.00293	0.00210	0.00048	0.00045	0.00340
Two-stage [(S5.2)(S5.7)(S5.8)]								
μ	0.90718	0.90716	0.90602	0.89558	0.90602	0.92119	0.92119	
σ^2	0.01877	0.01877	0.01868	0.01931	0.01696	0.01679	0.01155	
d	0.01181	0.01195	0.01204	0.01329	0.01204	0.00362	0.00376	
s.e.($\hat{\mu}$)	0.01149	0.01155	0.01169	0.01496	0.01169	0.00901	0.00901	0.00901
s.e.($\hat{\sigma}^2$)	0.00084	0.00084	0.00092	0.00127	0.00074	0.00072	0.00057	0.02404
s.e.($\hat{d}$)	0.00196	0.00199	0.03045	0.02915	0.03045	0.02537	0.00087	0.27337
Unbiased two-stage [(S5.2)(S5.11)(S5.12)]								
μ	0.90718	0.90716	0.90602	0.89558	0.90602	0.92195	0.92195	
σ^2	0.01877	0.01877	0.02122	0.02244	0.01895	0.01871	0.01244	
d	0.01181	0.01195	0.01390	0.01609	0.01390	0.00448	0.00467	
s.e.($\hat{\mu}$)	0.01149	0.01155	0.01257	0.01679	0.01257	0.00958	0.00958	0.00958
s.e.($\hat{\sigma}^2$)	0.00084	0.00084	0.00128	0.00199	0.00094	0.00092	0.00061	0.00172
s.e.($\hat{d}$)	0.00196	0.00199	0.00291	0.00447	0.00291	0.00101	0.00102	0.00634

because the denominator used in its calculation is, for ML, the total number of fetuses and, for REML, the same figure less one. For d, the difference is in terms of the cluster sizes (division by n_i or $n_i - 1$), which is more noticeable. All weighted estimators, except with equal weights, lead to very similar point estimates; this is in line with the simulation results. Even for equal weights, the difference is not worrisome. Proportional, equal, and approximate scalar weights are parameter-free and depend at most on the cluster size and/or the number of clusters per size. This explains why these estimators yield standard errors similar to the likelihood-based ones. Not surprisingly, because of their deviation from optimality, equal weights lead to increased uncertainty.

Table 3. NTP Data (EG). Cluster-by-cluster analysis. Maximum likelihood and weighted split-sample estimates (standard errors): (a) ML: maximum likelihood; (b) REML: restricted maximum likelihood; (c) Prop.: proportional weights; (d) Equal: equal weights; (e) Approx. sc.: like proportional weights, except that for b_k (6.11) is used; (f) Scalar: scalar weights, with the sub-sample specific weights plugged in for the parameters figuring in the weights; (g) Opt.: optimal weights, with the sub-sample specific weights plugged in for the parameters figuring in the weights. Proper: proper variances for optimal weights.

							Optimal	
Par.	ML	REML	Prop.	Equal	Approx. sc.	Scalar	Simpl.	Proper
Weighted [(S5.2)(S5.3)(S5.4)]								
μ	0.82952	0.82952	0.83342	0.84653	0.83342	0.84133	0.84133	
σ^2	0.00886	0.00886	0.00885	0.00899	0.00879	0.00878	0.00608	
d	0.01704	0.01724	0.01606	0.01536	0.01606	0.01381	0.01408	
s.e.($\hat{\mu}$)	0.01402	0.01410	0.01393	0.01485	0.01393	0.01346	0.01346	0.01346
s.e.($\hat{\sigma}^2$)	0.00041	0.00041	0.00046	0.00051	0.00044	0.00044	0.00031	0.00328
s.e.($\hat{d}$)	0.00265	0.00269	0.00264	0.00272	0.00264	0.00230	0.00230	0.00476
Two-stage [(S5.2)(S5.7)(S5.8)]								
μ	0.82952	0.82952	0.83342	0.84653	0.83342	0.84100	0.84100	
σ^2	0.00886	0.00886	0.00803	0.00814	0.00802	0.00802	0.00559	
d	0.01704	0.01724	0.01688	0.01621	0.01688	0.01476	0.01499	
s.e.($\hat{\mu}$)	0.01402	0.01410	0.01423	0.01522	0.01423	0.01379	0.01379	0.01379
s.e.($\hat{\sigma}^2$)	0.00041	0.00041	0.00037	0.00041	0.00037	0.00037	0.00029	0.03410
s.e.($\hat{d}$)	0.00265	0.00269	0.02814	0.02555	0.02814	0.02632	0.00243	0.05214
Unbiased two-stage [(S5.2)(S5.11)(S5.12)]								
μ	0.82952	0.82952	0.83342	0.84653	0.83342	0.83911	0.83911	
σ^2	0.00886	0.00886	0.00885	0.00899	0.00879	0.00878	0.00608	
d	0.01704	0.01724	0.01857	0.01833	0.01857	0.01657	0.01684	
s.e.($\hat{\mu}$)	0.01402	0.01410	0.01493	0.01665	0.01493	0.01452	0.01452	0.01452
s.e.($\hat{\sigma}^2$)	0.00041	0.00041	0.00046	0.00051	0.00044	0.00044	0.00031	0.00363
s.e.($\hat{d}$)	0.00265	0.00269	0.00302	0.00333	0.00302	0.00271	0.00271	0.00533

For the scalar and optimal estimators two issues need to be borne in mind. First, in principle they require knowledge of the true parameters. In the absence of them, plug-in estimates were used. Because of the independence between mean and variance parameters, both methods produce the same results for μ. Also, the estimates for μ are similar to the likelihood-based ones. For σ^2, this scalar-weight method works better than the optimal, matrix-based one. Because of their matrix nature, optimal weights are less stable when approximated. The standard errors are underestimated because uncertainty, stemming from plugging in the weights is ignored when using the 'simplified' precision estimates. When rectified ('proper' weights), there is no difference for the mean parameter, because

Table 4. NTP Data (DYME). Cluster-by-cluster analysis. Maximum likelihood and weighted split-sample estimates (standard errors): (a) ML: maximum likelihood; (b) REML: restricted maximum likelihood; (c) Prop.: proportional weights; (d) Equal: equal weights; (e) Approx. sc.: like proportional weights, except that for b_k (6.11) is used; (f) Scalar: scalar weights, with the sub-sample specific weights plugged in for the parameters figuring in the weights; (g) Opt.: optimal weights, with the sub-sample specific weights plugged in for the parameters figuring in the weights. Proper: proper variances for optimal weights.

							Optimal	
Par.	ML	REML	Prop.	Equal	Approx. sc.	Scalar	Simpl.	Proper
Weighted [(S5.2)(S5.3)(S5.4)]								
μ	0.84142	0.84141	0.84108	0.84861	0.84108	0.90166	0.90166	
σ^2	0.01031	0.01031	0.01072	0.01071	0.01034	0.01031	0.00700	
d	0.03657	0.03695	0.03102	0.03445	0.03102	0.00745	0.00755	
s.e.($\hat{\mu}$)	0.01926	0.01936	0.01780	0.02502	0.01780	0.01257	0.01257	0.01257
s.e.($\hat{\sigma}^2$)	0.00044	0.00044	0.00052	0.00079	0.00047	0.00046	0.00033	0.00308
s.e.($\hat{d}$)	0.00529	0.00537	0.00570	0.01043	0.00570	0.00159	0.00159	0.00329
Two-stage [(S5.2)(S5.7)(S5.8)]								
μ	0.84142	0.84141	0.84108	0.84861	0.84108	0.90009	0.90009	
σ^2	0.01031	0.01031	0.00975	0.00964	0.00945	0.00942	0.00650	
d	0.03657	0.03695	0.03199	0.03552	0.03199	0.00836	0.00845	
s.e.($\hat{\mu}$)	0.01926	0.01936	0.01804	0.02535	0.01804	0.01297	0.01297	0.01297
s.e.($\hat{\sigma}^2$)	0.00044	0.00044	0.00042	0.00059	0.00039	0.00039	0.00030	0.02568
s.e.($\hat{d}$)	0.00529	0.00537	0.03433	0.03113	0.03433	0.02215	0.00173	0.04036
Unbiased two-stage [(S5.2)(S5.11)(S5.12)]								
μ	0.84142	0.84141	0.84108	0.84861	0.84108	0.89672	0.89672	
σ^2	0.01031	0.01031	0.01072	0.01071	0.01034	0.01031	0.00700	
d	0.03657	0.03695	0.03690	0.04514	0.03690	0.01027	0.01037	
s.e.($\hat{\mu}$)	0.01926	0.01936	0.01937	0.02989	0.01937	0.01390	0.01390	0.01390
s.e.($\hat{\sigma}^2$)	0.00044	0.00044	0.00052	0.00079	0.00047	0.00046	0.00033	0.00353
s.e.($\hat{d}$)	0.00529	0.00537	0.00718	0.01542	0.00718	0.00205	0.00205	0.00382

the weights are parameter-free, but there is a strong difference for the variance components. Once the proper standard errors are calculated, it is clear that there is information loss because of using plug-in estimates in the weights, rather than the true ones.

8.2. Splitting by cluster size, with dose effect

While these results illustrate the explicit derivations with a constant mean, the data analysis in Section 8.1 does not do full justice to the actual design of the experiment, because the question of scientific interest is the dose-response relationship. Let x_i be the dose administered to cluster i, taking one out of 4 to

5 values. The dose levels for the DEHP study are given in Table 1. The model then is $\boldsymbol{Y}_i \sim N((\beta_0 + \beta_1 x_i)\mathbf{1}_n, \sigma^2 I_n + dJ_n)$. Because the mean and covariance parameters are functionally and statistically independent within a sub-sample of constant cluster size, the considerations presented for the constant-mean case will remain valid. The results of fitting this extended model to the DEHP, EG, and DYME compounds, under ML (and REML) on the one hand, and using split-sample methodology (with proportional, equal, and approximate scalar weights) on the other, are presented in Table 5. The results are comforting, showing that proportional and approximate scalar weights are a sensible choice. This is consistent with theoretical considerations, the simulations results, and the analysis in Section 8.1.

8.3. Cluster-by-cluster methods

We illustrate the cluster-by-cluster methods here. Results are presented in Tables 2–4, for DEHP, EG, and DYME, respectively. For brevity, attention here is confined to the case of no dose effect.

We consider three alternatives. In all three, (S5.2) is used for the mean. For the variance components, the pairs (S5.3)-(S5.4), (S5.7)-(S5.8), and (S5.11)-(S5.12) are used, respectively. Because these expressions are derived for a given cluster size, we need to supplement them with a weighting method. For comparison, the same choices are made as reported in Tables 2–4.

Even though the same estimator *per cluster size* is used for the mean in all three cases, the overall result is different for scalar and optimal weights because these depend on the estimated variance components. A relatively clear message is that proportional and approximate scalar weights show very good performance. This is pleasing, because these weights are parameter-free and hence easy to apply. As to which of the three versions is better is less clear, this differs somewhat from compound to compound and from parameter to parameter. All three show acceptable behavior. It is interesting to see that in some cases the cluster-by-cluster analysis is closer to ML than the analyses based on splitting per cluster size. Computationally, this approach allows for additional parallel processing, with all clusters analyzed in parallel and the results then combined.

9. Ramifications and Concluding Remarks

We considered the simple but insightful case of clustered data with a normal compound-symmetry structure and clusters of varying size. Here, there is no closed-form maximum likelihood estimator and maximization must proceed

Table 5. NTP Data (with dose effect). Splitting by cluster size. Maximum likelihood and weighted split-sample estimates (standard errors): (a) ML: maximum likelihood; (b) REML: restricted maximum likelihood; (c) Prop.: proportional weights; (d) Equal: equal weights; (e) Approx. sc.: like proportional weights, except that for b_k (6.11) is used.

Par.	ML	REML	Prop.	Equal	Approx. sc.
DEHP					
Interc. β_0	0.96986	0.96987	0.95982	0.95269	0.95982
Dose eff. β_1	−0.00077	−0.00077	−0.00042	−0.00029	−0.00042
σ^2	0.01876	0.01876	0.02122	0.02244	0.01895
d	0.00772	0.00792	0.00538	0.00508	0.00538
s.e.$(\hat{\beta}_0)$	0.01343	0.01357	0.01343	0.01609	0.01343
s.e.$(\hat{\beta}_1)$	0.00012	0.00012	0.00014	0.00018	0.00014
s.e.$(\hat{\sigma}^2)$	0.00084	0.00084	0.00128	0.00199	0.00094
s.e.$(\hat{d})$	0.00136	0.00141	0.00137	0.00204	0.00137
EG					
Interc. β_0	0.94228	0.94229	0.94654	0.95320	0.94654
Dose eff. β_1	−0.00009	−0.00009	−0.00010	−0.00010	−0.00010
σ^2	0.00879	0.00879	0.00847	0.00847	0.00833
d	0.00745	0.00765	0.00625	0.00593	0.00625
s.e.$(\hat{\beta}_0)$	0.01453	0.01470	0.01389	0.01406	0.01389
s.e.$(\hat{\beta}_1)$	0.00001	0.00001	0.00001	0.00001	0.00001
s.e.$(\hat{\sigma}^2)$	0.00041	0.00041	0.00044	0.00049	0.00042
s.e.$(\hat{d})$	0.00126	0.00130	0.00108	0.00107	0.00108
DYME					
Interc. β_0	1.01875	1.01876	1.02364	1.03680	1.02364
Dose eff. β_1	−0.00102	−0.00102	−0.00099	−0.00100	−0.00099
σ^2	0.01032	0.01032	0.01072	0.01071	0.01034
d	0.00795	0.00813	0.00581	0.00631	0.00581
s.e.$(\hat{\beta}_0)$	0.01356	0.01370	0.01335	0.02000	0.01335
s.e.$(\hat{\beta}_1)$	0.00006	0.00006	0.00006	0.00007	0.00006
s.e.$(\hat{\sigma}^2)$	0.00044	0.00044	0.00052	0.00079	0.00047
s.e.$(\hat{d})$	0.00126	0.00130	0.00110	0.00205	0.00110

iteratively. Moreover, there is no uniform optimal unbiased estimator and the MLE is only locally optimal.

When considering the collection of estimators obtained from analyzing the data for each cluster size separately, the MLE for the entire dataset is a vector linear combination of them, with the weights depending on the parameters. Based on theoretical results and simulations, as well as on data analysis, we found that equal weights and so-called approximately optimal weights do not perform well. Iterated optimal and proportional weights show excellent performance, and they

are simple and parameter-free. One refinement is that for the mean parameter and for the covariance term d weights should be chosen proportional to the number of clusters of a particular size, c_k, while for the measurement error variance σ^2 proportionality is to the product of the number of clusters of a given size and the cluster size, $c_k \cdot n_k$.

While most of our development is based on the simple, three-parameter compound-symmetry model, in the data analysis we considered a slightly expanded setting, in which the mean takes the form of a regression function. This suggests the use of our results in more elaborate settings, as long as some form of exchangeability prevails. One such setting is the meta-analytic evaluation of surrogate endpoints (Burzykowski, Molenberghs and Buyse (2005)), where two correlated endpoints rather than a single one are considered for each cluster (trial in this case). Admittedly, there may come a point where distinguishing between parameters where it is difficult to determine whether proportional weights or size proportional weights are to be preferred. Based on our simulation results, it may then be sensible to consider proportional weights for all parameters. In the case where clusters take the form of trials, the number of trials may be relatively small, and likely trial sizes are (almost) unique. Our split-sample method then implies that each trial is first analyzed separately, with overall estimates taking the form of linear combinations of trial-specific ones. To provide a formal basis for this, we considered the important special case of a cluster-by-cluster analysis. Such a method is consistent when the number of replicates per cluster (e.g., the number of patients per trial) increases more rapidly than the number of trials. Such an assumption is not realistic in the developmental toxicology setting considered in this paper, but may be sensible in a meta-analysis of clinical trials.

When clusters are very large, it may be attractive to further sub-divide them in sub-clusters. Such a splitting method was also considered by Molenberghs, Verbeke and Iddi (2011). Its use in our context would require further investigation.

In the NTP data, the observed cluster size is related to the dose applied. This suggests that it is useful to consider, at the same time, the impact of dose on the outcomes (e.g., fetal weight) as well as on cluster size. This brings us back to the informative cluster sizes mentioned in the Introduction. While work has been done in this area, it is of interest to combine the ideas developed in this paper with a model for cluster size.

Supplementary Materials

More detailed information can be found in thee accompanying Supplementary Materials. Section S1 explains the incompleteness in the compound-symmetry model based on the characterization of Hermans et al. (2017). The resulting lack of closed-form solutions for MLE are outlined in Section S2 and further calculations in Section S3. Background on the pseudo-likelihood-based split-sample method is given in Section S4. More on the derivation of weights for the compound-symmetry case are given in Section S5. Section S6 and S7 give more details about, respectively, a first and second simulation study. Section S8 describes the use of `R` for the analysis of the case study.

Acknowledgment

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.

References

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

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.

Arnold, B.C. and Strauss, D. (1991) Pseudolikelihood estimation: some examples. *Sankhya B* **53**, 233–243.

Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhya* **15**, 377–380.

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.

Burzykowski, T., Molenberghs, G. and Buyse, M. (2005). *The Evaluation of Surrogate Endpoints*. New York: Springer.

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.

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.

Fieuws, S., Verbeke, G., Boen, F. and Delecluse, C. (2006). High-dimensional multivariate mixed models for binary questionnaire data. *Applied Statistics* **55**, 1–12.

Follmann, D., Proschan, M. and Leifer, E. (2003). Multiple outputation: Inference for complex Clustered data by averaging analysis form independent data. *Biometrics* **59**, 420–429.

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*.

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

Laird, N. M. and Ware, J. H. (1982) Random effects models for longitudinal data. *Biometrics* **38**, 963–974.

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.

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.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

Neiswanger, W., Wang, C. and Xing, E. (2013). Asymptotically exact, embarrassingly parallel MCMC. arXiv preprint arXiv:1311.4780

Price, C. J., Kimmel, C. A., George, J. D. and Marr, M. C. (1987). The developmental toxicity of diethylene glycol dimethyl ether in mice. *Fundamental and Applied Toxicology* **8**, 115–126.

Price, C. J., Kimmel, C. A., Tyl, R. W. and Marr, M. C. (1985). The developmental toxicity of ethylene glycol in rats and mice. *Toxicology and Applied Pharmacology* **81**, 113–127.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I. and McCulloch, C. E. (2013). Bayes and big data: The consensus Monte Carlo algorithm. In: *Proceedings of he EFaBBayes 250 Conference* **16**.

Sikorska, K., Lesaffre, E., Groenen, P. F. J. and Eilers, P. H. C. (2013). GWAS on your notebook: fast semi-parallel linear and logistic regression for genome-wide association studies. *BMC Bioinformatics* **14**, 166.

Tyl, R. W., Price, C. J., Marr, M. C. and Kimmel, C. A. (1988). Developmental toxicity evaluation of dietary di(2-ethylhexyl)phthalate in Fischer 344 rats and CD-1 mice. *Fundamental and Applied Toxicology* **10**, 395–412.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* **21**, 5–42.

Van der Elst, W., Hermans, L., Verbeke, G., Kenward, M., Nassiri, V. and Molenberghs, G.

(2015). Unbalanced cluster sizes and rates of convergence in mixed-effects models for clustered data. *Journal of Statistical Computation and Simulation* **86**, 1–17.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

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.

I-BioStat, Universiteit Hasselt, Martelarenlaan 42, B-3500 Hasselt, Belgium.

E-mail: lisa.hermans@uhasselt.be

I-BioStat, KU Leuven, Kapucijnenvoer 35, B-3500 Leuven, Belgium.

E-mail: vahid.nassiri@kuleuven.be

I-BioStat, Universiteit Hasselt, Martelarenlaan 42, B-3500 Hasselt, Belgium.

I-BioStat, KU Leuven, Kapucijnenvoer 35, B-3500 Leuven, Belgium.

E-mail: geert.molenberghs@uhasselt.be

Manderville, The Wall, Ashkirk, Selkirk TD7 4NY, United Kingdom.

E-mail: mg.kenward@outlook.com

Janssen Pharmaceutica, B-2340 Beerse, Belgium.

E-mail: im.vanderelst@gmail.com

I-BioStat, Universiteit Hasselt, Martelarenlaan 42, B-3500 Hasselt, Belgium.

E-mail: marc.aerst@uhasselt.be

I-BioStat, KU Leuven, Kapucijnenvoer 35, B-3500 Leuven, Belgium.

I-BioStat, Universiteit Hasselt, Martelarenlaan 42, B-3500 Hasselt, Belgium.

E-mail: geert.verbeke@kuleuven.be

(Received January 2016; accepted April 2017)